

On the Blind Tasting of Wines: A New Method of Analysis and Beyond

Domenic V. Cicchetti, Ph.D.

Child Study Center and Departments of Psychiatry and Biometry
Yale University School of Medicine
New Haven, CT. 06510

Senior Research Scientist, Senior Biostatistician, and Senior
Research Psychologist

Yale Home Office
94 Linsley Lake Road
North Branford, CT 06471

dom.cicchetti@yale.edu

On the Blind Tasting of Wines: A New Method of Analysis and Beyond

Domenic V. Cicchetti, Ph.D.

In a recently published article in the journal *Chance*, Ashenfelter & Quandt (1999) re-analyzed the results of a Paris blind wine tasting that compared French Bordeaux and American cabernets in the heralded 1976 tasteoff that marked the 200th Anniversary or bicentennial celebration of the founding of America. Steven Spurrier, an Englishman and his American partner, Patricia Gallagher, hosted the event, and also served as two of the 11 judges. The remaining nine were notable French wine connoisseurs.

In a still more recent article that appeared in the New York Times in 2001, Frank Prial, an internationally famous wine connoisseur, again summarized the main findings. Both red and white wines were evaluated, but only the former will be re-analyzed here.

It was determined by Spurrier and Gallagher that a California cabernet was the winner, having defeated all the French challengers and this on French soil. These data were re-analyzed using ranks rather than mean ratings with the same conclusion being reached by Ashenfelter & Quandt (1999). The same conclusion was reached by the noted American wine critic, Frank Prial, as reported in 2001 in the New York Times.

This same conclusion is one that is held by most everyone in the wine consuming world, both to the chagrin of the French and to the jubilation of the Americans.

It is the purpose of the current research to do an in depth re-analysis of

the data to investigate whether the conclusion that has been reached is indeed valid. This will be accomplished by using state-of-the-art procedures; and to illuminate further the biostatistical and practical or clinical significance of the findings. The remaining sections of the report will describe in more detail: the wines in the taste-off; the tasters who evaluated them; the measurement instrument that was used to evaluate the wines; the design of the current investigation; and the results and their implications or heuristic value for further research in this intriguing and challenging area of scientific inquiry.

The Wines

There were 10 red wines in the tasting, 4 notable Bordeaux from France, and 6 notable cabernet sauvignons from California. The four French wines were a 1970 Mouton Rothschild, a 1970 Montrose, a 1970 Haut Brion, and a 1971 Leoville-Las-Cases from the Pomerol area of France. The California cabernets included a 1973 Stag's Leap, a 1971 Ridge Mt. Bello, a 1970 Heitz Cellars Martha Vineyards selection, a 1972 Clos du Val, a 1971 Mayacamas, and a 1969 cabernet from the Freemark Abbey winery.

The Tasters

The distinguished tasters were 11 in number. The 9 French tasters were: Odette Kahn, now deceased, who was then editor of the *Revue du Vin de France*; Jean-Claude Vrinat of the famous Restaurant Taillevent; Raymond Oliver of the restaurant Le Grand Vefour, now also deceased; the sommelier Christian Vanneque of the world renown Tour D'Argent; Aubere de Villaine of Domaine de la Romanee-Conti; Pierre Brejoux of the Institute of Appellations of

Origin; Pierre Tari of Chateau Giscour; Michel Dovaz of the Wine Institute of France; and Christian Millau, author of the famous and comprehensive guide for tourists. The remaining two, as previously noted, were the Englishman, Steve Spurrier, and his American partner Patricia Gallagher, who had both been impressed by their previous experiences with California wines, and who were responsible for organizing and hosting the event.

The Measuring Device

Wines were evaluated blindly and independently by each of the 11 judges, using a 20-point rating scale, such that increasingly higher scores indicated increasingly better quality of the wines being rated.

The Design of the Current Investigation

This study sought to answer a number of previously unaddressed questions:

1. Did the Americans really win the competition?
2. How reliable were the judges? This was accomplished by application of the intraclass correlation coefficient (Fleiss, 1981) utilizing a computer program described by Cicchetti & Showalter (1988).
3. Is it possible to discover a subset of tasters whose evaluations are notably more reliable than the remaining wine judges?
4. What advantages are there, if any, to converting scores from the 20-point measuring instrument to ranks?
5. How well or poorly do actual and ranked scores compare? and

6. What implications or heuristic value do the findings of this research have for further studies in this area of inquiry?

Results and Discussion

The first question is whether the previously reported results and their interpretation are indeed valid.

The data for each of the 10 wines are given in Table 1. Each wine is classified by Name (e.g., Mouton Rothschild, Stag's Leap); Origin (France, California); Mean Rating (1-20); Ranking; Standard Deviation of the Mean (SD); and Range of scores (lowest to highest).

Table 1. Scoring and Ranking of the Wines at the 1976 Paris Tasting

<i>Wine</i>	<i>Average</i>		<i>Standard</i>	
	<i>Scores</i>	<i>Rank (%)</i>	<i>Deviation</i>	<i>Range</i>
A. Stags Leap '73 (CA)	14.14	1(70.70)	1.70	10.0-16.5
B. Mouton '70 (FR)	14.09	2(70.45)	1.76	11.0-15.0
C. Montrose'70(FR)	13.64	3(68.20)	1.96	11.0-17.0
D. Haut Brion '70(FR)	13.23	4(66.15)	2.79	8.0-17.0
E. Ridge Mt. Bello '71(CA)	12.14	5(60.70)	3.44	7.0-17.0
F. Leoville-Las-Cases'71(FR)	11.18	6(55.90)	1.60	8.0-14.0
G. Heitz Martha's '70(CA)	10.36	7(51.80)	4.06	2.0-17.0
H. Clos du Val '72(CA)	10.14	8(50.70)	4.21	2.0-16.5
I. Mayacamus '71(CA)	9.77	9(48.85)	3.49	3.0-14.0
J. Freemark Abbey '69(CA)	9.64	10(48.20)	3.83	5.0-15.0

The data in Table 1 indicate clearly that although it is true that the California 1973 Stag's Leap cabernet sauvignon received both the "highest" average rating (14.14; SD = 1.70) and the "highest" ranking of 1, it is so closely followed by the French 1970 Mouton Rothschild (Average rating =14.09; SD=1.76), that all reasonable biostatisticians would be forced to agree that the two wines clearly ended in a statistical dead heat. To conclude otherwise would be to ignore a most fundamental of scientific facts, namely, the known chance variation around any statistic one chooses to derive. To make the point even stronger, one can easily convert the raw scores to percentages by multiplying by 5. This results in a 70.70% rating for the Stag's Leap cabernet and a totally indistinguishable rating of 70.45% for the Mouton Rothschild entry, hardly a meaningful difference. (Interestingly, the ranking method used by Ashenfelter & Quandt designates the third highest rated wine, the French Montrose, with a rating of 13.64 *higher* than the Mouton Rothschild with a higher average rating of 14.09! Such a data-distorting method lacks intuitive appeal).

It can also be seen, more generally that the remaining 3 French Wines were rated considerably higher than the remaining 5 California wines.

The ratings for these French wines ranged between 55.90% (Mean rating of 11.18), for the Leoville-Las-Cases pomerol and 68.20% (Mean rating of 13.64) for the Montrose Bordeaux entry. These 3 wines' ratings were ranked third, fourth, and sixth of the 10 wines that were evaluated in the tasting.

The ratings for the remaining American wines contrast sharply with these results. By ratings, the 5 remaining cabernets ranged between 48.20% (Mean rating of 9.64), for the Freemark Abbey cab to 60.70% (Mean rating of 12.14), for the Ridge Mt. Bello wine. The ordered rankings of these remaining American wines was 5,7,8,9, and 10.

These findings indicate clearly that the declaration that the Americans won the 1976 wine tasting is simply incorrect. It, in fact, gives little credence to the title of Frank Prial's (2001) New York Times article: "The day California shook the world."

Another rather serious problem with the Ashenfelter & Quandt (1999) analysis of the data is their weak attempt to assess the reliability of the ratings of the 10 wine entries by the 11 judges. This problem was obviated in this re-analysis by applying the well-known intraclass correlation coefficient, R_i (e.g., Fleiss, 1981) to the data, using a computer program reported previously by Cicchetti, Aivano, & Vitale (1976) and later, in an updated and more comprehensive version by Cicchetti & Showalter (1988). The model of the R_i that was applied assumes that the same set of judges made all of the ratings, as was of course true for the wine tasting data discussed here.

Results indicated the following: The overall agreement level, corrected for chance, produced an R_i of only 0.22. Although statistically significant at a probability level (or p value) of $<.01$, this value, by the criteria of both Cicchetti & Sparrow (1981), and Fleiss (1981) can be considered poor in a practical or clinical sense.

The next question posed is whether it is possible to identify from among the 11 tasters a subset whose wine ratings are distinctively more reliable than the remaining tasters? The answer is yes, as I shall now demonstrate.

After the overall R_i analysis was completed, the program calculated separate R_i values between the 10 judges, on a pair-by-pair basis. This produced a total of $(10 \times 9) / 2 = 45$ R_i coefficients. These were then rank ordered from highest to lowest, thereby providing all the information required to distinguish the 5 most reliable tasters (designated numbers 2,5,6,7 and 11) from the remaining or 6 least reliable judges (1,3,4,8,9 and 10).

The 11 judges were assigned the following IDs, as derived from the information provided in Ashenfelter & Quandt (1999, p.18): Brejoux=1; de Villaine=2; Dovas=3; Gallagher=4; Kahn=5; Millau=6; Oliver=7; Spurrier=8; Tari=9; Vanneque=10; and Vrinat=11.

The same type of R_i analysis that was used for determining the reliability of all 11 tasters was applied to the two subsets of judges. As expected, the R_i value for the 5 most reliable subset of tasters ($R_i=.69$) and considered "Good" by the aforementioned criteria of Cicchetti & Sparrow (1981), was considerably higher than the corresponding R_i value for those 6 judges previously designated as the less reliable ones. In this case, $R_i=.31$, which is considered "Poor" by the same criteria. This strategy allows for follow-up of these two groups of tasters to determine the consistency of their levels of reliability in future tastings.

The next question I sought to answer is what possible advantages are there in using actual scores for rating wines as opposed to converting these

scores to ranks, as was undertaken by Ashenfelter & Quandt (1999)? These two authors are adamant about the necessity to convert the raw scores (1-20) to ranks, in order to minimize the presumed stronger and relatively disproportionate influence that any one judge might have on the results by using only a restricted range of the scale (say, 19 or 20) to rate the wines rather than a broader spectrum of scores within the full scale range of 1-20. What the authors failed to do is to consider whether this legitimate concern actually affected the ratings. Since this is an empirical question, I checked the data to determine whether such a potentially biasing phenomenon was, in fact, operating. This was accomplished by simply observing the actual range of scale points that was assigned by the full panel of 11 judges to each of the 10 wine entries. The results are both clear and highly informative. With the 10 wines rank-ordered by the size of the mean or average rating, as A, B, C, D, E, F, G, H, I and J, respectively, as in Table 1, the following corresponding scoring ranges are as shown here: Wine A, 10.0-16.5; Wine B, 11-15; Wine C, 11-17; Wine D, 8-17; Wine E, 7-17; Wine F, 8-14; Wine G, 2-17; Wine H, 2-16.5; Wine I, 3-14; and Wine J, 5-15. Finally, the range of scores that was applied across all 10 wines by all 11 tasters was between a low of 2 and a high of 17. These data indicate overwhelmingly that there was no wine bias operating, of the genre feared by Ashenfelter and Quandt (1999). Moreover, I might argue that if indeed one of the judges applied the scale using only 2 scale points, say 19 and 20, I would want to examine this further, by questioning her/him after the tasting was completed. Such a taster, for example, might have believed in a tasting similar to the one reported here that there really was no

basic difference in the overall quality of the 10 wines, that, in fact, they were all, in his opinion, of the very highest quality. Given the outstanding selection of wines that comprised this group of 10, this opinion certainly has a ring of plausibility. A third reason for rejecting the idea of using ranks in lieu of the actual ratings is that the ranks completely hide or mask the underlying scores they represent. Why should the same score be ranked differently as a function of different raters. Why, in fact, construct the 20 point scale in the first place? And, finally, the conversion of raw scores to ranks necessitates the application of nonparametric statistics, which are known by biostatisticians to be less powerful than their parametric counterparts; and, finally, the p values or results that are obtained using nonparametric rather than parametric statistics are, except under very unusual circumstances, very similar in size. A striking example of this derives from an empirical look at another statement made by Ashenfelter and Quandt (1999) to justify their conversion of scores to ranks: "As the table indicates, there is a loose agreement between the ranking of the wines using the average grade and the average rank" (p.19).

Nothing could be further from the truth. The actual correlation between average ranks and average scale scores is a whopping .93.

As a second example, Ashenfelter and Quandt applied Kendall's coefficient of concordance, a statistic that simply measures judges' agreement between the rank orderings, again ignoring and masking completely the raw data from which they derive. However, despite this distortion, the value of 0.24 that

was obtained compares quite favorably to the 0.22 obtained by applying the more appropriate R_i that was utilized here.

Finally, there is the issue of what to do with a Kendall's Tau coefficient once it is obtained. It cannot be interpreted much beyond the fact that higher values indicate higher agreement in rank orderings of the data. In distinct contrast, the R_i measures the true level of chance-corrected *agreement* based upon the actual rather than the distorted data. This is especially important in situations such as wine tasting when all that we really have to rely upon is the reliability of ratings. To convert to ranks can very well eliminate the influence of one or more judges who may be the "correct" judge, in the sense of having, as one example, a more stringent and well-defined set of scoring criteria than her/his more consistent tasters who share in common a lack of such well-defined criteria.

And so, on the topic of whether to use or not use ranking as a primary data analytic strategy in future wine tastings, I would for all the presented arguments much prefer that the original scale scores be used instead of the ranks. Absent this, I would at the very least opt for both procedures, parametric and nonparametric to be employed so those proponents of parametric approaches can see what the results look like before the original data are distorted.

The final issue to be addressed here is what implications or heuristic value might the results of this research have for future studies of wine tasting.

A number of questions that need to be answered come to mind: First, would the most reliable tasters established in a given tasting continue as such, or are these initial findings just another chance phenomenon? Was Steven Spurrier correct, as quoted by Ashenfelter & Quandt (1999, p. 17), when he said, referring to the very 1976 experience that “the results of a blind tasting cannot be predicted and will not even be reproduced the next day by the same panel tasting the same wines.” If replicable, how are they affected by changes in the types of wines? Can the less reliable tasters be taught to become more reliable? Can a more reliable wine rating scale be constructed? Numerous other critical questions can be posed and answered about the vagaries and vicissitudes of the wine tasting experience, including how the suggested paradigm would work with wine aficionados, such as the current author, who are not at the same level and sophistication as that august group of 11.

References

- Ashenfelter, O, and Quandt, R. (1999). Analyzing a wine tasting statistically. Chance, 12, 16-20.
- Cicchetti, D.V., Aivano, S.L. & Vitale, J. (1976). A computer program for assessing the reliability and systematic bias of individual measurements. Educational and Psychological Measurement, 36, 761-764.
- Cicchetti, D.V., & Sparrow, S.S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. American Journal of Mental Deficiency, 86, 127-137.
- Cicchetti, D.V., & Showalter, D. (1988). A computer program for determining the reliability of dimensionally scaled data when the numbers and specific sets of examiners may vary at each assessment. Educational and Psychological Measurement, 48, 717-720.
- Fleiss, J.L. (1981). Statistical Methods for rates and proportions. New York, NY: Wiley (2nd ed).
- Prial, F. (2001). The day California shook the world. New York Times, May 9th.