

## Documented Disagreement Among Wine Experts: A Rational Response for Consumers

Dom Cicchetti, Ph.D.

Department of Biometry  
Yale University of Medicine  
New Haven, CT 06520

**Objective:** The purpose of this presentation is three fold: First, to document the levels of agreement among three notable sources of wine raters, namely, Robert Parker, Jancis Robinson, and the Wine Spectator (notably, James Suckling) on the evaluation of the multiple tastings of the 2004 Bordeaux; and second, to make rational sense of the expected disparities in ratings of any given wine, with the target of concern being the wine consumer. Another way of understanding this issue is to think in terms of how best the consumer might behave in the context of appreciable differences in wine preferences among the oenological elite.

**Methodology:** The data base was derived from web-site <http://www.bordoverview.com> that lists a comprehensive array of Bordeaux wines rated by Robert Parker (RP), Jancis Robinson (JR); and the Wine Spectator (WS) and by other lesser known wine experts. The data are further classified according to: Location (e.g., Haut-Medoc, St-Emilion, Pauillac; St-Estephe); Left Bank vs. Right Bank; Wine Classification (1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup>, 5<sup>th</sup> Growth); and year of release; 2004, 2005, 2006, 2007, and 2008).

**Focus:** All Bordeaux wines rated in 2004, by; Robert Parker, Jancis Robison, and the Wine Spectator (James Suckling's ratings).

**Wine Rating Scales:** The scales utilized by RP and JR are both conceptually and structurally very similar. On the other hand, the wine rating scale utilized by JR, while conceptually similar to the other two, is dissimilar in structure. Both RP and WS utilize a scale ranging between 50 and 100, in which: < 70%= Poor, or Below Average; 70%-79%=Fair= 80%-89%=Good; and 90%-100%=Excellent. Although JR uses a scale ranging between 12 and 20 points, they translate into categories that are equivalent or conceptually similar to the RP and WS scales such that: 12-13=Below Average; 14-15=Average, or Fair;16-17=Above Average (or Good); and 19-20=Excellent,

**Data Analytic Strategies:** To compare RP with WS; and RP with JR in terms of overall agreement, corrected for chance. The statistic utilized in the first comparison was the Intraclass Correlation Coefficient (ICC), due to John Bartko (1966). It can range between -1 and +1, in the manner of the standard Pearson Product Moment Correlation. Cicchetti & Sparrow (1981) classified levels of ICC, in terms of practical or clinical significance, as follows: <0.40=Poor; 0.40-0.59=Fair; 0.60-0.74=Good; and 0.75 and Above = Excellent chance =corrected levels of agreement.

Also, in the case of two ratings, the level of agreement before chance correction can be calculated directly as  $(ICC + 1)/2$ , as demonstrated many decades ago by

Robinson (1957). Also, because of the mathematical similarity between Kappa, Weighted Kappa and the ICC, the criteria used to define levels of practical or clinical significance of ICC values can also apply to levels of Kappa or Weighted kappa.

**Results:** These are divided into two comparison categories: RP vs Ws; and RP vs. JR.

*A. RP vs. WS:*

134 2004 Bordeaux wines were rated independently by RP and WS. The ICC value was 0.52, that was statistically significant at  $p = 0.001$ ; with a level of practical or clinical significance of Fair or Average agreement. Applying the aforementioned formula due to Robinson (1957), an ICC value of 0.52 translates into an agreement level, before chance-correction, of  $1.52/2=76\%$  (also Fair agreement, by the criteria of Cicchetti, Volkmar, Klin, & Showalter (1995).

*B. RP vs. JR:*

184 wines were tasted by RP and JR. Since RP and JR used different scales, it became necessary to equilibrate the two scales, to result in conceptually comparable instruments.

Recall that RP utilizes a wine rating scale ranging between 50 and 100. However, this scale abbreviates conceptually into a 4 category scale whereby:

0=< 70%= Poor (Unacceptable); 1=70%-79%= Fair (Average); 2=80%-89%=Above average; and 4=90%-100%=Excellent

Similarly, JR utilizes a scale ranging between 12 and 20, 2whereby:

0=12-13=Unbalanced (Unacceptable); 1=14-15=Average; 2=16-17=Above Average; and 3=18-20=Excellent

*C. Reliability statistic:* Weighted Kappa, utilizing a scoring system developed by Cicchetti & Sparrow (1981) in which:

Ratings in complete agreement (0-0; 1-1; 2-2; and 3-3) receive an agreement weight of 1; 1-2; 2-1; 2-3; and 3-2 pairings receive a weight of 0.80; 0-1 and 1-0 ratings receive a weight of 0.60; pairings of 1-3 and 3-1 receive a weight of 0.40; pairings of 0-2 and 2-0 are assigned a weight of 0.20; and pairings as far apart as possible, 3-0 and 0-3, receive a weight of 0.

Application of Weighted Kappa (Cohen, 1968; Fleiss, Cohen, & Everitt, 1968) resulted in 83% agreement, compared to 81% expected by chance alone. This resulted in a Weighted Kappa value of only 0.12, or very Poor chance-corrected agreement, by the criteria of Cicchetti & Sparrow (1981) and Fleiss (1981).

Further analysis revealed that of the 114 of 184 wines in which there was less than perfect agreement, in 110 or 97.3% of them, RP assigned a higher wine rating than did JR.

If there had been no such bias, one would expect that when there was a disagreement, JR would be as likely as RP to rate the wine of lesser or better quality. The excess of 47.3% represented the bias reflecting RP's consistently higher wine ratings than those of JR. This was statistically significant at  $p < 0.001$ , by application of McNemar's chi-square(d) statistic (McNemar (1945)).

RP rated 2 wines as Above average or of Good quality that JR rated as Unacceptable

RP rated 19 wines as Excellent that JR rated as of only Average quality;

For another 38 wines, RP rated them as Above average, while JR classified them as of Average quality and

RP rated 51 wines as Excellent that JR rated as being only Above average.

**Discussion:** Conducted for the first time, this analysis of agreement levels in the rating of close to 200 2004 Bordeaux shows that while RP and WS (James Suckling) agree at a Fair or Average level of agreement; RP and JR disagree considerably; their levels of agreement are very Poor by conventional biostatistical standards; and this is reflected in JR's tendency to consistently rate the great majority of these wines (62%) as being of higher quality than did JR.

Given the published and personally documented differences between RP and JR in their published tasting notes of the same wines, these results, though not previously documented, are not surprising.

In attempting to make sense of the disparity in wine ratings among the putative experts, one might again heed the cogent advice of Jancis Robinson(1997) when she stated that:

“Individual wine consumers are better off, my argument goes, following an individual wine critic's preferences and prejudices and getting to know how they relate to their own—in the same way that we filter what we're told by, say, individual theatre or film critics. For, make no mistake about it, wine judging is every bit as subjective as the judging of any art form” (p. 319).

Putting this sage advice into a wider context, there are no right or wrong answers to stated wine preferences. Some of us like fruit-bombs; some go for more structure, subtlety, and balance. And, yes, still others like both styles of wine, depending upon the context in which they are consumed.

In the final analysis, it is perhaps fair to say that it is heterogeneity in wine tasting preferences that makes the oenological world go 'round!

## References

- Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3-11.
- 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., Volkmar, F., Klin, A., & Showalter, D. (1995). Diagnosing autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. *Child Neuropsychology*, 1, 26-37.
- Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31, 651-659.
- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. New York, NY: Wiley (2<sup>nd</sup> ed.).
- Fleiss, J.L., Cohen, J., & Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323-327.
- Robinson, J. (1997). *Tasting Pleasure; Confessions of a Wine Lover*, New York, NY: Penguin.
- Robinson, W.S. (1957). The statistical measurement of agreement. *American Sociological Review*, 22, 17-25.