Cicchetti, Domenic V.; Cicchetti, Arnie

Article

Two enological titans rate the 2009 Bordeaux wines

Wine Economics and Policy

Provided in Cooperation with:
UniCeSV - Centro Universitario di Ricerca per lo Sviluppo Competitivo del Settore Vitivinicolo, University of Florence

Suggested Citation: Cicchetti, Domenic V.; Cicchetti, Arnie (2014) : Two enological titans rate the 2009 Bordeaux wines, Wine Economics and Policy, ISSN 2212-9774, Elsevier, Amsterdam, Vol. 3, Iss. 1, pp. 28-36, https://doi.org/10.1016/j.wep.2014.01.001

Available online at www.sciencedirect.com

ScienceDirect

Wine Economics and Policy 3 (2014) 28–36

www.elsevier.com/locate/wep

Two enological titans rate the 2009 Bordeaux wines

Dom Cicchetti[a,*], Arnie Cicchetti[b]

[a]Department of Biometry, Yale University School of Medicine New Haven, CT 06511, United States
[b]Certified Specialist of Wine (CSW), San Anselmo, CA 94960, United States

Received 18 September 2013; received in revised form 18 December 2013; accepted 23 January 2014
Available online 12 February 2014

Abstract

The purpose of this research was to compare the ratings of 237 2009 Bordeaux wines by Jancis Robinson (JR) and Robert Parker (RP). Results indicate that the level of agreement was 81.77% as compared to an expected agreement of 80.92%. This produced a chance-corrected agreement level of only 4%. Though statistically significant at a level of 0.02, the practical or clinical usefulness of such a result was essentially nil. Further analyses shed light on the phenomenon in that: there was complete agreement on only 27% of the wines; and on the 73% or 172 wines upon which there was a disagreement, RP scored all of them higher than did JR or: 7 wines as GOOD that JR scored as FAIR; 22 wines as Excellent that JR rated as Fair; and 143 wines as Excellent that JR rated as Good. Finally, the authors provide preliminary evidence that expert wine tasters appear to fall into two distinct sub-groupings, here designated as A and B. While the tasters within each group agree substantially with one another; the tasters in Group A disagree substantially with the tasters in Group B. The implications for future enological research are discussed.

Keywords: Wine experts; Statistical and clinical significance

1. Introduction, background and aims

Recently, Jancis Robinson (JR) and Robert Parker (RP) rated the 2004 Bordeaux vintage wines (Cicchetti and Cicchetti, 2013). This vintage was a problematic one sandwiched between the two more highly acclaimed Bordeaux vintage years of 2003 and 2005. When corrected for chance, the agreement level was only 12%. Though statistically significant, this result is devoid of any practical or clinical usefulness. By this, we mean that the result occurred beyond chance expectation, but that it does not reach the minimal level of practical usefulness, or a chance-corrected level of agreement of at least 0.40 (Cicchetti and Sparrow, 1981). It underscores the fact that given a large enough number of cases, almost any level of agreement will be statistically significant, while not being of any practical value.

To make this point more explicit, one of the French Chardonnay wines in the heralded 1976 Paris wine tasting was the Clos des Mouches, with a chance-corrected agreement level, among the 11 judges, of 0.10 or 10%, a result that did not even approach statistical significance; nor was it of any practical significance. With 15 raters this same level of 0.10 shows a higher level of statistical significance (a so-called trend in the right direction); while increasing the number of wine judges to 20 now produces a result that becomes statistically significant at a probability level of 0.03, while still being of little practical or clinical significance at the low chance-corrected level of 0.10. This has been referred to as the "big *N* phenomenon" (Cicchetti, 2007).

1.1. Wine critics comment on Bordeaux 2009

The authors ask, somewhat whimsically, is the 2009 Bordeaux Vintage perhaps flirting with schizophrenia? This

*Corresponding author. Tel.: +1 203 376 2913; fax: +1 293 488 4218.
E-mail addresses: dom.cicchetti@yale.edu (D. Cicchetti), acicchetti@earthlink.net (A. Cicchetti).
Peer review under responsibility of Wine Economics and Policy.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.wep.2014.01.001

question is posed with due restraint while pouring over books, blogs and pod-casts, with experienced wine tasters commenting, as follows.

Robert Parker (2010) states that the 2009 vintage, "may turn out to be the finest vintage I have tasted in 32 years of covering Bordeaux." Fine, but the title of his article is "Robert Parker's 2009 Ratings Out: Beware of Bordeaux Frauds." How does one make sense of this proclamation?

Steven Spurrier (2010) of Decanter Magazine and co-host of the famous 1976 French and American Bordeaux taste-off, begins with the laudatory statement that the 2009 wines "are great", while ending with the contradictory call "for a restraint in pricing".

Jean-Charles Cazes (2010) (Lynch-Bages, Les Ormes de Pez) is quoted by James Suckling, Wine Spectator, as saying' "The wines are rich and powerful, yet smooth and refined at the same time, and the acidity is good." Well which are they- rich/powerful or smooth/refined? And good acidity? Again, a mixed message to the enologic community.

Will Lyons (2010), in his Wall Street Journal, decries that in the case of the Left Bank wine makers, "…perhaps the best wines they have ever made." This is certainly a most powerful statement. And finally, Eric Arnold, Contributor to Forbes magazine, says in his 2010 blog "2009 Bordeaux is great? Pass."

So what is all the fuss about? If this is such a great vintage or as some bloggers have pronounced, "The Vintage of the Century," why not have astronomical pricing? Look at the '05's- well that's the rub. It seems that some collectors like Eric Arnold bought into the 2005 hype and paid handsomely for them. He says, "I bought into the hype early on, and am the proud owner of a mini-fridge full of 2005 wines that are worth less than I paid for them." Mr. Arnold says that he won't get fooled again. He'll be purchasing the '08's, because to him, they are, "the best deal in town." One can understand his frustration. Already in this century the 2000s, 03's, and '05's have earned critical acclaim and near perfect scoring from the wine critics. How can the 2009 vintage really be any better? But when Christian Moueix of Petrus' fame is quoted as stating about the 2009 vintage, "I'm usually very critical, but it's maybe my best vintage in my 40 years of experience," it's hard not to get caught up in the hype.

1.2. The 2009 Bordeaux vintage: weather and the quality of wine

Bordeaux grapes in the beginning of 2009 faced a cold winter with average temperatures below those experienced in 2008. This forced the vineyards into dormancy and the ability to prune on time. Nature took its course. The cold snap continued into March and budding did not take place until early April. Everything thus far followed textbook viticulture.

However, as the experience of vineyard managers reminds us, a great early season does not necessarily make a great wine in the fall. As projected, Merlot had bud break earlier than Cabernet. This pattern would follow throughout the vintage, which of course is normal.

In May, a hailstorm broke in the Right Bank hurling hail the size of small snowballs. It was devastating to structures, autos and vineyards. It hit above and to the North of St. Emilion, where vines suffered mass devastation. Actually, if one looks at a map of Bordeaux, part of the "upper" region like Pomerol is to the west of St. Emilion, not north. Thus the authors want to emphasize that the storm hit due north, not to the west. It is a critical geographical issue. June and July saw pleasant days that helped in the ripening of fruit. August was warmer and the sugar content grew. Soon September and October came, bringing more of the same warm weather. Grapes hung with ample sugars, but the purple hues had not overtaken the green fruit color. Would vineyards have over-ripe fruit without the coloring matching the sugar levels? Vineyard managers could be seen tasting grapes in the vineyards to watch in trepidation, should sugar levels become too high before the grapes matured.

Early October saw vineyard hands picking Merlot that had high sugar levels, although later the must would prove to have good acid and tannic levels. The crop seemed in balance, and later picking of Cabernet would also hold true. Overall, the best of the vintage seemed to have the hoped for consistency that wineries embrace. The first growths had wines with such power and intensity that it led some to deem them as perhaps the best wines that they had ever made.

But there is a caveat. Will these wines age well over time, or will they falter? Even Decanter Magazine cautions that these wines may age poorly. Is it true, or a mea culpa for the 2005 vintage? Ah that vintage – the one where Chateau and Estate Wines, one of the premier marketing companies and a wing of Diagio, was found to be dumping a large portion of the '05's on the open market. Is that a harbinger of things to come for the '09? Thus the hand wringing and consternation on the part of the wine critics.

In the words of Chris Kissack (2009), Ph.D., the Wine Doctor, "there is more to a great vintage than a few great wines. The 2009 vintage is a story of inconsistency." To Mr. Kissack there are just too many weird, extracted and overly alcoholic wines in this vintage. His overall evaluation goes as follows: "Let me state this clearly: 2009 – talking specifically of the red wines from Bordeaux – is *not* a great vintage."

One logical conclusion that can be drawn from this and the aforementioned array of differing evaluations by some of the leading wine authorities is that the 2009 Bordeaux may be a great Bordeaux vintage – but then again…

While this section has dealt with the weather and its effect upon the quality of wine, it is important, in the context of this enological narrative to compare this phenomenon with the extent to which the ratings of wine critics affect the quality of wine, as raised by one of the reviewers.

1.3. How do wine critics' ratings correlate with the quality of wine?

A number of critically important papers have addressed this issue. These results are summarized and masterfully integrated in a recent article by Karl Storchmann (2012) who cites Ashenfelter's

Bordeaux equation which predicts the price of wine on the basis of the following variables: age of the vintage; average temperature over the growing season (April–September); amount of rain in August; and average temperature in September. As Storchmann notes, the Bordeaux equation: "shows that wine experts are less accurate than quantitative methods in predicting a wine's quality. Because Bordeaux wines are not ready to be consumed before an age of about 8–10 years, vintage assessments need to forecast a vintage's quality. Although the Bordeaux equation's predictions with an R^2 of 0.838, an R correlational equivalent of 0.92 (for vintages spanning between 1952 and 1980-authors' insertion) are fairly accurate, experts steadily adjust their ratings as more information about a wine's drinkability becomes available. Particularly mediocre vintages are often rated too highly" (ibid, p. 12). Thus, while the 1975 Bordeaux vintage was rated as mediocre by the Ashenfelter Bordeaux equation, Parker awarded it 95 points. He was then forced to down-grade the wines, as they matured, and then recommended that they be consumed quickly, rather than stored for 8–10 years, as would be undertaken for a wine meriting a score of 95 out of a possible 100 points.

Not to be outdone, Parker referred to Ashenfelter's (1997, 2008) Bordeaux equation using the following rather colorful phraseology, "really a Neanderthal way of looking at wine. It is so absurd as to be laughable", in short "an absolute sham" (Storchmann, 2012, p. 10). Similarly Ayers (2007) quoted Robert Parker's evaluation of the author of the Bordeaux equation, as follows: Ashenfelter is "rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director." However, as Ayers noted, the Ashenfelter equation correctly forecasted that the 1989 Bordeaux vintage would be excellent, and the 1990 better yet. Contrast this to Parker's evaluation: "In 1989, he awarded this very vintage only 88 points and recommended that the wines be consumed immediately, rather than stored" (Storchmann, 2012, p. 10). In commenting on this phenomenon, Ayers (2007) put it quite forcefully, when he wrote "Take that, Robert Parker."

All this said, it is instructive that weather variables are substantially better predictors of wine quality than is expert judgment (Ashenfelter, 1995; 2008; and Storchmann, 2012). To this point, it is well worth noting another important reviewer observation, namely, that the effect of experts' wine ratings upon wine pricing appears to be minimal, at best. Thus, Ali et al. (2007) found that Robert Parker's ratings had an effect equal to less than 3 Euros per bottle for 2003 Bordeaux wines. In a more recent study of the blind tasting of South African (SA) wines, the results were even more negative. The correlations (Pearsonian Rs) between wine tasting preferences and wine pricing, for SA Sauvignon Blanc and SA Pinotage wines were, respectively, only 0.06 and 0.15, values that are just slightly above zero (Cicchetti, under review).

2. Data and methodology

2.1. The Bordeaux 2009 data base

The data base (http://www.bordoverview.com) consisted of 237 2009 Bordeaux wines that were evaluated independently by both JR and RP using wine rating scales that are structurally different, conceptually quite similar, and translatable one into the other, as will next be demonstrated. (For a comparison of extant wine rating scales, the interested reader is referred to Cicchetti and Cicchetti, 2009.)

2.2. Jancis Robinson's wine rating scale

JR employs a 12–20 point scale whereby:

12–13 Represents below average wine quality.
14–15 Denotes average quality.
16–17 Informs that the wine is above average quality.
18–20 Tells the consumer that the wine has been judged to be of superior quality.

2.3. Robert Parker's wine rating scale

RP applies a 50–100% wine rating scale in which:

$< 70\%$ = Below average wine quality
70–79% = Average quality
80–89% = Above average quality and
90–100% = Superior wine quality

2.4. "Equating" the JR and RP wine rating scales: pros and cons

The translation of the JR scale into that of RP is facilitated by the fact that the maximum JR score of 20, when multiplied by 5, produces the maximum JR wine score of 100. The intermediate scores also fall into place, as will now be shown (Table 1).

However, as one reviewer correctly observes, a JR rating of 12, that she characterizes as a "faulty or Unbalanced" wine, is analogous to an RP rating of 60–69, one he would describe as "a below average wine containing noticeable deficiencies, such as excessive acidity and/or tannin, an absence of flavor, or possibly dirty aromas or flavors".

At the other extreme, what JR describes as a "truly exceptional" wine, would be analogous to RP's more verbose depiction of "an extraordinary wine of profound and complex character displaying all the attributes expected of a classic wine of its variety. Wines of this caliber are worth a special effort to find, purchase, and consume."

Table 1
Converting JR into RP wine rating scale.

JR scale	RP scale	Qualitative evaluation
12–13	60–65%	Below average or poor
14–15	70–75%	Average or fair
16–17	80–85%	Above average or good
18–20	90–100%	Superior or excellent

Similarly, the reviewer observes that JR's rating of 19 classifies such a wine as "a humdinger" while the more prolix RP's analogous rating of 90–95 describes "an outstanding wine of exceptional complexity and character. In short, these are terrific". No argument here. JR and RP are on the same "purple page" to use JR's words in a more limited enological context.

But how about the remaining JR and RP descriptors? How do they match up, one to the other? Here, as the same reviewer would aver, it is not so straight forward. For example, JR's rating of 17 merits a "superior" rating by JR and "a very good wine displaying various degrees of finesse and flavor as well as character with no noticeable flaws," by RP; while what JR describes as a "deadly dull" wine, deserving a low score 14, has as its RP analog, a rating of 70–79 that is described as "an average wine with little distinction except that it is soundly made. In essence, a straightforward, innocuous wine". Clearly, what appears as a run-of-the-mill wine to RP is experienced as "deadly dull" by JR.

Extending the reviewer's concern further, let us consider the more ideal case, one in which the wine tasters use essentially the same rating scale, as in the case of comparing wine evaluations of Robert Parker, James Suckling (representing the Wine Spectator) and Steven Tanzer. It is doubtful that the same score (say 85) has the same connotative meaning for this trio of putative wine experts. So while it is not ideal, the conversion of JR's ratings to those of RP, by a multiplier of 5 is probably the very best one can accomplish, under the circumstances.

2.5. Evaluating levels of agreement between JR and RP

A biostatistically valid method of determining how any two independent examiners agree with each other needs to focus consensus at three levels: the Percentage of Observed agreement (PO); the Percentage of Chance agreement (PC) and the Percentage of Chance-Corrected agreement or (PO-PC)/(100-PC), (Cohen, 1960). Each of these will be discussed in turn.

2.6. Calculating PO and partial rater agreement weights

The level of agreement between JR and RP is based upon a weighting system developed by Cicchetti (1976). In an earlier publication, Hall (1974) argued convincingly that linear partial agreement weights should be used when the Weighted Kappa statistic is applied to establish levels of inter-rater agreement, as in the present investigation. Hall correctly noted that other systems, notably one based upon quadratic partial weights, will weight equally probable rater errors quite unequally. Cicchetti (1976) supported this position, but also demonstrated, for the first time, that there are, in fact, two distinct types of ordinal scales that can be designated as either continuous-ordinal (CO) or dichotomous-ordinal (DO). Put simply, the CO scale measures entities that can only be described in terms of what one might call degrees of presence (e.g., anxiety, with criteria for defining "slight", "moderate" or "severe").

Clinically speaking, a rater's disagreement between "slight" and "moderate" is no more serious than one between "moderate" and "severe" anxiety. Contrast this with the dichotomous-ordinal (DO) rating scale, one which ranges from "absence" to varying degrees of "presence", such as "slight", "moderate" and "severe." What characterizes or typifies this scale is that, unlike the aforementioned CO rating scale, any rater disagreement that confuses "presence" of a given entity with its "absence" will per force receive less "credit" or "weight" than one the same number of scale points apart that does not confuse "absence" with "presence". The bio-behavioral sciences abound with DO clinical rating scales. As one example, consider criteria for examining a psychotic patient for hallucinatory symptomatology rated as "none", "slight", "moderate" or "severe". Here the discrepancy between "none" and "slight" must, per force, be considered more clinically serious than disagreements between either "slight" and "moderate" or between "moderate" and "severe" – this despite the fact that each of these three rater disagreements is the same one category apart. Put succinctly, it becomes clear that confusing the "absence" of hallucinatory behavior with "slight" hallucinatory behavior is clearly more clinically serious than confusing either "slight" with "moderate" or "moderate" with "severe" hallucinatory symptomatology. The general assumption here is that holding constant the number of categories separating a pair of ratings, confusing "presence" with "absence" of a disorder is more serious than confusing degrees of "presence" of that disorder.

Reasoning beyond the distinction between CO and DO rating scales, a further question that arises is: how does one determine for a given CO or DO scale, the actual linear weighting system that would be necessitated? The general principle would be that the weighting system for both the CO and the DO rating scales would range between: 100% (for each rater pairing in complete agreement); and 0% (for each rater pairing that is maximally apart). And, finally, intermediate partial agreement weights would be assigned, following the same principle, that is, the agreement weights would decrease, in a linear fashion, the greater the number of ordinal categories separating each given pair of ratings.

As also given in Cicchetti (1976), the formula for determining the number of weights for any CO rating scale (beginning with complete agreement and ending with complete disagreement) is given by the below formula, in which k refers to the number of ordinal categories:

$$k-1/k-1,\quad k-2/k-1,\quad k-3,\ldots\ldots k-k/k-1 \tag{1}$$

Thus, for a 6 category CO ordinal rating scale, the respective weights would be: 5/5 = 100% for 1–1, 2–2, 3–3, 4–4, 5–5, and 6–6 pairings; 4/5 = 80% for 1–2, 2–1, 2–3, 3–2, 3–4, 4–3, 4–5, 5–4, 5–6, and 6–5 pairings; 3/5 = 60% for 1–3, 3–1, 2–4, 4–2, 3–5, 5–3, 4–6, and 6–4 pairings; 2/5 = 40% for 1–4, 4–1, 2–5, 5–2, 3–6, and 6–3 pairings; 1/5 = 20% for 1–5, 5–1, 2–6, and 6–2 pairings; and 0/5 = 0% for 1–6 and 6–1 rater pairings. The linear weights for DO rating scales take into account that disagreement pairings that are a given number of points apart, and involve presence–absence discrepancies merit less partial credit than disagreement pairings the same number of points apart that do not involve "presence"–"absence" discrepancies. This, again, follows the principle that confusing "presence"

with "absence" is more serious, all things considered, than confusing degrees of "presence".

The number of agreement weights (W) that applies to a given DO scale, is simply the number of points on the scale added to the number of partial agreement categories. This means a 5 category DO scale would have $5+3=8$ weights. In a more formal manner, the number of DO weights can also be determined by the simple formula

$$W = 2(k-1) \quad (2)$$

where k once again refers to the number of ordinal scale points. Thus, a three category DO scale will have 4 weights (1, 0.6667, 0.3333, and 0); a four category DO scale has 6 weights (1, 0.80, 0.60, 0.40, 0.20, and 0); and so on (Cicchetti, 1976).

Flashing forward to the weighting schemata applied to assess levels of agreement between JR and RP in the rating of 237 2009 Bordeaux wines, the following emerges. There are four ordinal categories representing the quality of a given wine: 0=poor/unacceptable (<70%); 1=Fair/Average (70–79%); 2=good/above average (80–89%); and 3=excellent/superior (90-100%). Because of the "unacceptable" (poor)–"acceptable" (fair, good, excellent) dichotomy, the scale defines itself as DO, with $2(k-1)$ or 6 partial agreement weights, emerging as: 100% (0–0, 1–1, 2–2, 3–3); 80% (1–2, 2–1, 2–3, 3–2); 60% (0–1, 1–0); 40% (1–3, 3–1); 20% (0–2, 2–0); and 0% (0–3, 3–0).

Pertaining to the ideas expressed in this paper, the scale used to compare the wine ratings of JR and RP would be classified as Dichotomous-Ordinal (DO). This special type of rank-ordered or ordinal scale has two qualities vis a' vis its substantive meaning. The dichotomous aspect has as its defining feature that the scale differentiates first between wines that are unacceptable (scores below 70%) and acceptable (scores between 70% and 100%). The ordinal aspect of the scale denotes the fact that there are three rank orderable levels of wine Acceptability, namely, Average, Above Average, or Excellent. The aforementioned Cicchetti and Sparrow (1981) paper reported chance-corrected inter-examiner pairings (Here JR vs RP wine ratings), where P=POOR; F=FAIR; G=GOOD; and E=Excellent wine quality, as follows: Table 2.

Table 2
Rater agreement weights for assessment of reliability.

JR–RP scorings	% Agreement
Poor–poor	100
Fair–fair	100
Good–good	100
Excellent–excellent	100
Fair–good	80
Good–fair	80
Good–excellent	80
Excellent–good	80
Poor–fair	60
Fair–poor	60
Fair–excellent	40
Excellent–fair	40
Poor–good	20
Good–poor	20
Poor–excellent	0
Excellent–poor	0

2.7. Calculating PC

The level of agreement expected on the basis of chance alone follows the same law of chance-probability (PC) that we learned in Statistics 101. Recall that if there were 90% red balls and 10% white balls in each of two jars and we wished to know what percentage of red ball and white ball matches would occur if we randomly selected the balls, pair-by-pair, from each of the 2 jars (without replacement), we would obtain $(0.9\times0.9)=0.81=81\%$ agreement on the red balls and $(0.1\times0.1)=0.01=1\%$ agreement on the white balls, or 82% overall agreement by chance alone.

Now the math is the same for paired wine ratings, the only added issue, is that as we did for PO, we need to account for levels of partial chance agreement.

2.8. Chance-corrected rater agreement or (PO-PC)/(100-PC)

The concept of chance-corrected agreement was introduced by Jacob Cohen in 1960 in the context of his widely known and widely applied Kappa statistic. The related statistic for rank-ordered data is the Weighted Kappa statistic (Cohen, 1968).

Note from the formula that Kappa, (and, in fact, Weighted Kappa) means, in words, that one is comparing the difference between observed and chance agreement, as compared to the maximum difference that can possibly occur, the latter being 100% minus the level of chance agreement.

2.9. Determining statistical significance of chance-corrected agreement

To determine level of statistical significance, the Kappa or Weighted Kappa value is simply divided by its standard error. This produces the familiar Z score that is then referred to a Table of Z values to determine the probability (p) or odds that the Kappa or Weighted Kappa value occurred by chance alone, When Z reaches a level of 1.96 it indicates that the probability that such a Kappa or Weighted Kappa value occurred by chance alone is only 5%, meaning we can be 95% certain that the result did not occur by chance. This is the familiar gold standard that we apply as the holy grail of interpreting the results of our research endeavors, be they enological or otherwise derived.

2.10. Statistical significance does not infer clinical significance

The problem in interpreting the meaning of say, an enological result that is statistically significant, is that with a large enough sample size, even the most meaningless of results will be statistically significant, Thus a correlation of only 0.05 will be statistically significant at the aforementioned biostatistically sanctioned level of chance occurrence of 5% when based upon a sample size of 1000. This same logic applies to

Kappa, Weighted Kappa and the results of statistical tests in general.

2.11. Determining levels of practical or clinical significance

The criteria we shall apply to determine whether the chance-corrected level of agreement between JR and RP is of any practical or clinically meaningful value were developed by Cicchetti and Sparrow (1981), deriving from the previous work of Fleiss (1981) and Landis and Koch (1977). They are, as follows: Table 3.

Table 3
Guidelines for interpreting Kappa values.

Kappa/Weighted Kappa	Clinical significance
<0.40	Poor
0.40–0.59	Fair
0.60–0.74	Good
0.75–1.00	Excellent

2.12. Extent of bias in JR and RP ratings

The last question that needs to be discussed here is to what extent are JR and RP biased in their wine ratings? This simple concept rests upon the biostatistical fact that whenever there is less than perfect agreement on given wine ratings, by chance alone we would expect RP to score as many wines higher than JR as she would score higher than would he.

As an example, if JR and RP disagreed in their scorings on 100 of the 237 wines, then we would expect JR to give higher scores than RP on 50 of these wines and that RP would give higher ratings than JR on the remaining 50. This would signal that there was no bias in the disagreed upon wine ratings.The chi-squared statistic that was developed by McNemar (1947) will test the amount of bias, as compared to the 50–50% or no bias level.

3. Results

In all, JR and RP rated 237 Bordeaux wines of the 2009 vintage. In applying the statistics just described, the following results were obtained: the overall level of agreement (PO) was 81.77%. The level of agreement expected by chance was 80.92% and the level of chance-corrected agreement (PO-PC)/(100-PC) was 0.04, or only 4%. Because it was based upon a large sample size ($N=237$ wines), the result was statistically significant at the 0.02 level of probability. However, a Weighted Kappa value of only 0.04 is closer to 0, than to the minimal value of 0.40 required for a level of fair or average chance-corrected agreement (Cicchetti and Sparrow, 1981).

To understand this result better, it is important to refer to an early, classic, and seminal paper by Robinson (1957), who focused upon a measure of chance-corrected inter-rater agreement that is applicable to both ordinal and interval variables, namely, the intra-class correlation coefficient (R_i). Robinson noted that the R_i has the same range as the standard Product Moment Correlation Coefficient (R), or between -1 and $+1$. Robinson was the first to show that in the case of 2 raters, there is a mathematical relationship between Rater Agreement (A) and R_i, such that: $A=(1+R_i)/2$, so that for an R_i, of 0.40, the lowest level of the CS criteria considered clinically meaningful, Agreement becomes 70%. To continue with the history of the problem, Cohen developed Kappa in 1960, and Weighted Kappa in 1968. Fleiss (1975) demonstrated the mathematical equivalence between Kappa for dichotomous variables and the R_i. Earlier, Fleiss and Cohen (1973) showed a mathematical equivalence between Weighted Kappa and the R_i. And because of these equivalencies, we have a veritable family of inter-rater reliability statistics.

Now, if we flash forward to the results of this investigation, the Weighted Kappa value of just 0.04, translates into an agreement level of only (1.04)/2, or 52%, a quite dismal result. In order to understand in more depth the meaning of the exceedingly low level of agreement between JR and RP, we shall examine next the aforementioned level of inter-rater bias.

JR and RP were in agreement on only 65/237, or 27% of their ratings. This means they were in disagreement on the remaining $172/237=73\%$ of the wines. For each of these 172 wines, RP gave a higher score than did JR! It is truly remarkable that there was not a single wine that JR rated higher than RP, when they were in disagreement. Application of the McNemar (1947) test of bias, for correlated proportions, was statistically significant at $p<0.0001$.

4. Discussion and conclusions

The next section of this report will be organized around two fundamental research questions: first, what seems to underlie the varying levels of disagreement between expert judges in their evaluation of the same wines? and second, what qualifies one as an enological titan?

4.1. Hypothesizing about the why's of disagreement among experienced wine critics

There are some preliminary data from two sources that are consistent with the hypothesis that there may be two distinct groups of professional wine tasters, who display strikingly dissimilar taste patterns. Designating the groups as A and B, the authors make a tri-partite hypothesis, such that: first, the A tasters would agree appreciably with each other; and, similarly, the B tasters would also be in high agreement with one another; however, the tasters in Group A would be in substantial disagreement with the tasters in Group B. There is currently some preliminary support for such an enological scenario.

The first source of information derives from the disparate wine evaluations of the highly visible 2003 Premier Grand Cru Pavie St. Emilion from the Bordeaux Right Bank. The tasting notes and corresponding scores, by six enological experts (Liberman, 2004) have been reported, as the following:

As we read between the enological lines (Table 4) the following conclusions begin to emerge:

Table 4
Comparing tasting notes for premier grand Cru Pavie 2003.

Wine expert	Score (%)
Jancis Robinson (JR) "Completely unappetizing overripe aromas. Why? Porty sweet. Oh, REALLY! Port is best from the Douro, not St. Emilion. Ridiculous wine more reminiscent of a late-harvest Zinfandel than a red Bordeaux with its unappetizing green notes."	60
Michael Broadbent (MB) "Very deep, extraordinary nose, slightly fishy, tarry; fairly sweet, full-bodied, powerful, dense, and again tarry."	70
Clive Coates (CC) "Anyone who thinks this is good wine needs a brain and palate transplant. This wine will be scored simply as undrinkable."	Undrinkable
Robert Parker (RP) "An off-the-chart effort…a wine of sublime richness, minerality, delineation, and nobleness…Inky/purple, to the rim, it offers up provocative minerals, black and red fruits, balsamic vinegar, licorice and smoke. It traverses the palate with extraordinary richness as well as remarkable freshness and definition. The finish is tannic, but the wine's low acidity and higher than normal alcohol (13.5 percent) suggests it will be approachable in 4–5 years…A brilliant effort, it, along with Ausone and Petrus, is one of the three great offerings of the Right Bank in 2003."	96–100
James Suckling (JS): Wine Spectator "Super-ripe and almost jammy. Very New World on the nose, but impressive; Bordeaux on the palate. Berries, raspberries and strawberries; hint of wood. Full-bodied with ripe and sound tannins and a long finish. Chewy. Got to like this."	95–100
Stephen Tanzer (ST) "Aromas of cassis, violets, minerals, and licorice; thick on entry, then chewy as a solid on the mid-palate, with powerful, super-saturated dark berry and mineral flavors and enough ripe acidity to give the wine shape and freshness. Best today on the great building finish, which features huge but thoroughly ripe tannins and palate-saturating berry and mineral flavors. An impressively rich, structural wine that wears its high alcohol gracefully."	92–95

1. There is impressively high agreement among the 6 experts that the Premier Grand Cru is a very full-bodied, fruit-forward wine, in fact, a wine that could easily be described as a veritable "fruit-bomb."
2. The six wine tasters divide rather easily into two camps, with JR, MB, and CC forming what can be labeled Group A tasters; and RP, JS, and ST comprising the Group B tasters.
3. It is also clear from the ratings that JR, MB, and CC are in substantial agreement that the 2003 Pavie is of rather poor quality; whereas the remaining tasters are in high agreement that the 2003 Pavie is of excellent to superior quality.

Consistent with these findings, a comprehensive analysis of average Bordeaux ratings, based upon 399 of the 2004–2008 Bordeaux wines, showed remarkably similar average ratings by RP, ST, and the Wine Spectator (James Suckling); these were, respectively: 91, 90, and 91. Similarly, the triads of average scores for the three tasters, classified by first, second, third, fourth, and fifth growths, were, as follows: Table 5.

Table 5
Ratings of Bordeaux first, second, third, fourth, and fifth growths.

Growth	Robert Parker	Stephan Tanzer	Wine Spectator
1	96	94	95
2	92	91	92
3	90	89	90
4	89	88	89
5	90	89	90

These average ratings are very similar. In fact, such minor differences can be expected to occur in any test retest tastings of replicate wines by the same taster.

While these two sets of data support the authors' aforementioned tri-partite hypothesis, they cannot be viewed as definitive, because of differences in: sample sizes (a single wine vs 399 wines), data analytic strategies; Bordeaux vintage years and varietals. This means that future research will be required to resolve these fundamental problems. The authors are in the process of designing additional enological research to address these critical issues.

Despite what remains to be accomplished, before more definitive conclusions can be drawn, one consistent finding that remains indisputable is the known vast differences in the rating of wine quality by putative wine experts. This presents as a potentially serious economic issue, as it impacts upon the wine producer, sales person and wine consumer. How, in effect, does one make sense of widely disparate wine ratings? An impressively articulated and well-reasoned solution that seems designed to untie this veritable Gordian knot has been provided in the writings of Thompson et al. (2008), who conclude that the differences in wine evaluations among the experts: "…suggest that consumers look for a rater whose tastes correspond with their own, and then put more credence on the ratings from that source…" (ibid, p. 6).

In this way, consumers begin to build a repertoire of wines that have the imprimatur of that expert, or group of experts –

with whom they agree most closely. It is also important to note that this advice to the consumer is very consistent with that of the aforementioned and highly acclaimed British wine expert Jancis Robinson (1997), who classifies the rating of wine an art form, such as ratings of films, musical productions, or other forms of theatrical performance. In attempting to conceptualize further, one begins to experience the complexity involved in assessing levels of consistency in the evaluation of wines among putative experts. Referring again to the 2003 Pavie, there is agreement among all tasters that the wine is redolent of fruit. However, the ratings divide into two rather distinct subgroupings depending primarily upon whether the tasters do or do not like fruit-bombs. Note that the scores that are given are based precisely upon this preference or lack thereof. And, finally, if we were to base levels of agreement upon tasting notes, as illustrated in Table 4, the extent of disagreement would reach across all of the expert tastings. In fact, such rater differences would seem to dwarf any that might have been caused by the equating of JR's wine rating scale (the multiplicand) to that of RP by the multiplier of 5 points.

4.2. What makes one an enological titan?

In closing, the authors focus upon another critical question pertinent to this research endeavor: A reviewer asks, what in fact makes one a wine expert or enological titan? This is a difficult and complex question to answer fully, vis a' vis JR and RP. Jancis Robinson is a Master of Wine (Wine-Searcher Staff, 2013). In fact, she rose to enological fame, following her becoming the first Master of Wine outside the wine trade. Robert Parker, on the other hand, has no formal training in wine tasting (Langewiesche, 2000). He is, in this respect, a self-made, albeit world famous, wine critic, who's enological career was launched when he predicted the later acclaimed greatness of the 1982 first growth Bordeaux wines. A reigning and well respected enologist in the field, Robert Finigan, spoke negatively about this vintage, describing it as "overly-alcoholic" (Steiman, 2011). As history was the judge, Parker has been proven "right" and Finigan "wrong". Finigan's fame was diminished considerably and Parker's enological fame skyrocketed, and to apply an old adage, "the rest is history".

All this said, experts and non-experts alike tend to agree that the acid test of the ability to evaluate wines successfully is to be able to identify them correctly in a blind wine tasting. To this point, Langewische (2000) writes that Parker told him "in a matter-of-fact way that he remembers every wine he has tasted over the past 32 years and, within a few points, every score he has given as well". Parker was given the opportunity to test his claim in a blind wine tasting of 15 of the heralded 2005 Bordeaux wines that he had evaluated, 2 years previously – in 2007 – and scored as his "favorite wines of the vintage." The wines and their respective Parker scores were: Angelus (98), Cos d'Estournel (98), Ducru Beaucaillou (97), Haut Brion (97), Lafite Rothschild (96+), La Mission Haut Brion (97), Larcis Ducasse (98), Latour (96+), L'Eglise Clinet (100), Margaux (98+), Montrose (95), Pape Clement (98), Pavie (98), Le Gay (95), and Troplong Mondot (99). As Dr. Vino (2009, p. 5) described the results: "…Parker upended the order of his published ratings of the wines and, in the process, could not correctly identify any of these wines. In print, he awarded L'Eglise Clinet, a Pomerol, a score of 100 points. While he did call it his second favorite wine of the night, it is interesting to note that he could not pick out this wine in the lineup (he thought the actual L'Eglise to be Cos, a wine that is not only from across the river, but from St. Estephe, an appellation known for the extreme tannic structure of the wines). In that same vein, he mistook Lafite, a Pauillac, for Troplong-Mondot, a new wave St. Emilion. Dr. Vino (ibid) concludes that "blind tasting can be ruthless in its outcomes." How much of this phenomenon is due to changes in a wine over a 2-year period (wine is a living organism) remains an unknown.

References

Ali, H.H., Lecocq, S., Visser, M., 2007. The impact of gurus: Parker and en primeur wine prices. J. Wine Econ. (Working paper).

Arnold, E., 2010. Commentary: Forbes Magazine, April 1.

Ashenfelter, O., 1997. A new objective ranking of the chateaux of Bordeaux. Liq. Assets 13, 1–6.

Ashenfelter, O., 2008. Predicting the prices and quality of Bordeaux wines. Econ. J. 118, 40–53.

Ayers, I. 2007. How computers routed the expetrts- A commentary by Ian Ayers' 86. Financial Times. August 31.

Cazes, J.-C., 2010. Commentary: Wine Spectator Magazine, March 19.

Cicchetti, D.V., 1976. Assessing inter-rater reliability for rating scales: Resolving some basic issues. Br. J. Psychiatry 129, 452–456.

Cicchetti, D.V., 2007. Assessing the reliability of blind wine tasting. J. Wine Econ. 2, 196–202.

Cicchetti, D.V. Blind tasting of South African wines: a tale of two methodologies (Under review).

Cicchetti, D.V., Cicchetti, A.F., 2009. Wine rating scales: assessing their utility for producers, consumers, and oenologic researchers. Int. J. Wine Res. 1, 73–83.

Cicchetti, D.V., Cicchetti, A.F., 2013. As wine experts disagree, consumers' taste buds flourish: the 2004 Bordeaux. J. Wine Res. 24, 311–317.

Cicchetti, D.V., Sparrow, S.S., 1981. Developing criteria for establishing inter-rater reliability of specific items: applications to assessment of adaptive behavior. Am. J. Ment. Defic. 86, 127–137.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 23, 37–46.

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220.

Fleiss, J.L., Cohen, J., 1973. The equivalence of weighted kappa and the intraclass correlation as measures of reliability. Educ Psychol. Meas. 33, 613–619.

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, 2nd ed. Wiley, New York.

Hall, J.N., 1974. Inter-rater reliability of ward rating scales. Br. J. Psychiatry 125, 248–255.

Kissack, C., 2009. Commentary on Bordeaux 2009: First report. Wine Doc.

Landis, J.R.,, Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics. 33, 159–179.

Langewische, W., 2000. The million dollar nose. Atl. Mon. 286, 42–70.

Liberman, M., 2004. Grand cru smackdown, 1–3, June 2nd.

Lyons, W., 2010. Commentary: Wall Street J., April 29th.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157.

Moueix, C., 2009. Commentary: 2009 Wine Experience: Right Bank Bordeaux tasting, Wine Spectator, October 27th.

Parker, R., 2010. Commentary: Robert Parker's ratings out: beware of Bordeaux frauds. Wine Advocate, October 27th.
Robinson, J., 1997. Tasting Pleasure; Confessions of a Wine Lover. Penguin, New York.
Robinson, W.S., 1957. The statistical measurement of agreement. Am. Sociol. Rev. 22, 17–25.
Spurrier, S.., 2010. Commentary: Decanter News, April 10th .
Steiman, H., 2011. Remembering Bob Finigan. Wine Spectator, October 5th.
Storchmann, K., 2012. Wine economics. J. Wine Econ. 7, 1–33.
Suckling, J., 2010. 2009 Bordeaux barrel tasting. Wine Spectator, March 10th.
Thompson, G.M., Mutkowski, S.A., Bae, Y., Ielacqua, L., Oh, S.B., 2008. An analysis of Bordeaux wine ratings, 1970-2005: Implications for the existing classifications of the Medoc and Graves, 2008. Cornell Hosp. Rep. 8, 4–18.