AMERICAN ASSOCIATION OF WINE ECONOMISTS

AAWE WORKING PAPER
No. 164
Economics

Blind Tasting of South African Wines: A Tale of Two Methodologies

Domenic Cicchetti

www.wine-economics.org

Aug 2014
ISSN 2166-9112

Blind Tasting of South African Wines: A Tale of Two Methodologies

Dom Cicchetti, Ph.D.
Department of Biometry
Yale University School of Medicine
New Haven, CT 06520

Keywords: Sauvignon Blanc; Pinotage; replicate tastings; wine rankings; wine scorings

Abstract

The author compared two methods for assessing the reliability of blind wine tastings of South African Sauvignon Blanc and Pinotage white and red wines. The first method required the taster to rank each of the 8 white and red wines from 1 for the most preferred to 8 for the least preferred, with tied ranks permitted when 2 or more wines were equally preferred. The second used scores representing a taster's perceived quality of wine as: 50%-69% =Poor/Unacceptable; 70%-79%= Fair/Mediocre; 80%-89%= Good/Above Average; and 90%-100%= Excellent to Superior. Results indicated that the two methods provided quite different results with respect to the critical issue of identifying replicate wines; while the identification of replicate white and red wines proved successful when wine scores were used, not a single wine taster could successfully identify replicate white or red wines when using ranking methods. This finding, plus the fact that very different scorings can produce the very same rankings, such as 1, 2, and 3 being given scores of 60, 64, 66 (Poor/Unacceptable); 70,72, 76 (Fair/Mediocre); 80, 83, 87 (Good/Above Average; and 90, 95, 98. (Excellent to Superior), led to the conclusion that such flaws in the ranking method render it of very limited usefulness in the evaluation of the results of blind wine tasting.

-

Purpose

The purpose of this research is to evaluate two methods for analyzing the results of a blind wine tasting, vis a` vis: the correct identification of randomly placed replicate wines; wine preferences; the extent to which the price of a wine is related to level of preference; and the overall levels of reliability of the wine tasters.

The Wines

Each of the 16 wines was produced in South Africa. The varietal of each of the 8 white wines was Sauvignon Blanc; while the varietal of each of the 8 red wines was Pinotage.

The Tasters

The 15 tasters were randomly drawn from among the participants of the seventh annual meeting of the American Association of Wine Economists (AAWE) held in Stellenbosch, South Africa. The same 15 tasters evaluated both the white (Sauvignon Blanc) and red (Pinotage) wines. The characteristics of each white and red wine are depicted in Tables 1 and 2, namely: vintner, vintage year, alcohol level, and retail cost in American currency.

- Insert Tables 1 and 2 about here-

How the Wines Were Evaluated

Each judge was asked in each of the two wine tastings to use two methods to evaluate the wines. . The first was to rank each of the 8 wines in the two competitions from 1 to 8 with 1 indicating the most preferred and 8 the least preferred wine. If two or more wines were equally preferred, then they would receive the same rank. In the case of an odd number of ties, the tied rank would be the middle value, so that if the ties spanned positions 5, 6, and 7, the tied rank for all three positions would be 6. The more cumbersome way to arrive at this same value would be to add the three numbers and divide by the same number 3, as in (5+6+7)/3=6. The procedures are the same when the ties span an even number of positions, as when there is a tie across ranks 5, 6, 7, and 8. Here, the simple way to determine the tied rank is to take the average of the middle 2 ranks. In this example (6+7)/2 = 6.5. The cumbersome/unnecessary way of arriving at the same result is,

again to sum the ranks and divide by the number of them to produce (5+6+7+8)/4=6.5.

The second method of wine evaluation involved applying a Wine Spectator rating scale to express the perceived quality of a given wine on a scale ranging between 50% and 100%, by which: 50-69=Poor/Unacceptable; 70-79=Fair/Mediocre; 80-89=Good/Above Average; and 90-100=Excellent to Superior.

In order to screen for the reliability of the tasters, three wines were replicates of the same white or red wines, and they were served from the same bottle in a randomized order.

Results

White Tasting of Sauvignon Blanc

Reliability of Replicate White Wines

In order to address this question appropriately, the enological researcher is obligated to set meaningful criteria to define what shall constitute successful or unsuccessful evidence of replication of a tasting result. For ranks, the criterion was the taster had to be within one rank for the three replicates. Not a single taster met criterion. These data are highlighted in Table 3.

The criterion for replicability of the scoring of both white and red wines derives, logically, from the aforementioned number/range of points that define a particular wine quality as Poor/ Unacceptable (<70%); Fair/Mediocre (70-79); Good/Above Average (80-89); or Excellent to Superior (90-100). These ranges span between 10 and 11points. To reflect this numerical reality, the replicability range was set at $\pm$ 5 points. This criterion was honored whether or not the range crossed perceived quality rating categories such as replicate scores of, say, 75 (Poor); 80 (Good), and 85 (Good), that would be scored as a successful triplicate replication, in the same manner as scores of, say, 70, 75, and 79, each of which would be classified as Fair/Mediocre.

When this criterion was applied to test replicability of Sauvignon Blanc tastings, the test proved successful across the average scorings of the 15 tasters,

with scores of 77, 82, and 82. Further analysis revealed that 10 of the 15 tasters met the replicability criterion, as well.

-Insert Table 3 about here-

Reliability of Replicate Red Wines

As was true for white wines, none of the 15 tasters met the replicability criterion of within one rank ordering.

Application of the replicability criterion for the scoring of the red wines was, again, successful. Here the three replicate average ratings were: 75, 76, and 78, spanning a very narrow range of only 3 points. Consistent with the white wine results, 10 of the 15 tasters successfully met the $\pm$ 5 point criterion.

Overall Taster Reliability for White Wines

The statistic of choice here is the appropriate model of the intraclass correlation coefficient (ICC)-e.g., as noted by Fleiss, Levin, & Cho-Paik (2003). This reliability statistic is part of a family of mathematically equivalent procedures, for nominal/dichotomous data (Fleiss, 1975); and ordinal and interval data, as demonstrated by Fleiss & Cohen (1973). The model used here applies when judges are considered the major focus of the analysis, and the same set of judges/tasters evaluates each case (or wines, in our situation).

These statistics can be interpreted as either significant, i.e., at or beyond the traditional 0.05 level of probability; and in terms of its level of clinical significance.

Application of the ICC indicated a chance-corrected agreement level of 0.3147 or 0.31, which, though statistically significant, does not meet criterion for clinical or practical significance by the criteria of Cicchetti & Sparrow (1981) or by those of Fleiss, Levin, & Cho Paik (2003), whereby ICC values below 0.40 are considered Poor.

Since the application of ranks resulted in a completely unsuccessful replication of the same white and red wines poured from the same bottle, it became unnecessary to consider the method further. Another way of stating this is that the ranks-as opposed to wine scores- produced inaccurate data, as reflected in the failure to replicate as occurred when the actual scores were utilized.

Overall Taster Reliability for Red Wine

Application of the ICC (e.g., Bartko, 1976; Shrout & Fleiss, 1979) across the 15 tasters produced a near zero value of 0.09 that was neither statistically nor clinically meaningful. Again, since the application of ranks to test for the success of replicability failed completely, it indicated that to consider ranks further would not be a meaningful pursuit.

The Relationship Between Wine Preference and Wine Cost

There was no statistically or clinically significant correlation between wine scores and the cost of the wine. This proved true for both the Sauvignon Blanc wines (Pearson R=0.06), or approaching zero; and similarly, a correlation of only 0.15 for the Pinotage wines.

Discussion

Several basic issues need to be discussed here. The first is to attempt to investigate why the agreement levels on both the Sauvignon and Pinotage were so low. To gain a better understanding of this phenomenon, we need to examine the extent of the variability of wine scorings on a wine by wine basis. A second critical issue is to understand further what type of information is produced by using ranks, rather than scores, in the analysis of the results of blind wine tastings. And a third issue is to discuss the heuristic value of this enological research. Put simply, what hypotheses or research questions derive from the results of this investigation?

While the *average* ratings varied between the narrow band of 77 and 82, across the 8 white wines, the extensive variability, as measured by the range is given, first for the white wines; then for the reds.

On a wine by wine basis, the range of scores for each of the 8 white wines was, as follows: 70-95; 58-97; 61-90; 50-90; 72-90; 69-91; 50-95; and 55-97.This wide range of scores for each of the 8 white wines is between 50 (Poor/Unacceptable/Lowest Score Possible) and 95 (Excellent to Superior). These enormously wide ranges of scores for each of the 8 white wines are masked, or become invisible if the focus is solely upon the aforementioned narrow range of average wine ratings.

The results of the ranges of scores for red wines is consonant with these findings. As was true of the scoring of white wines, the average ratings across the 8 wines ranged between a narrow band of 72 and 78. However, on a wine by wine basis, the ranges for the 8 red wines, were, respectively: 60-90; 55-89; 60-92; 60-90; 65-90; 60-87; 60-89; and 60-95. This wide range of scores is between 55 (Poor/Unacceptable) and 95 (Superior). What these data indicate is that the average scores, once again, hide the very wide range of scores on each of the 8 wines.

These widely discrepant results are reminiscent of the rating of the 2003 Grand Cru St. Emilion Bordeaux wine by prominent and respected wine gurus, namely: Jancis Robinson, Michael Broadbent, Steven Tanzer, James Suckling of the Wine Spectator, and Robert Parker. Their respective percentage ratings of the quality of this controversial wine were, as follows: 60, 70, 94, 98, and 98. While the tasting notes indicated that the wine was perceived by each expert as very full-bodied, such that it merited the descriptor "fruit-bomb", the wide variation in perceived quality of the wine ranged between Poor (60) and Superior (98). It is rather well known in the enological world that both Jancis Robinson and Michael Broadbent abhor "fruit-bombs", while the remaining wine experts,-who might be deemed Los Tres Amigos de Vino, simply adore "fruit-bombs." Hence, the extremely low level of inter-taster agreement. It is instructive that our representative sample of AAWE participants at the 2013 Stellenbosch blind wine tastings showed very similar results, as reflected in widely varying palate preferences.

The next issue pertains to the consequence of using ranks instead of scores to evaluate blind wine tasting results. What does each of the two methods offer vis a` vis each other?

Well, it depends upon what question one wishes to answer. If the only question is the relative ranking of 2 or more wines, specifically, which of the wines does one most prefer, quite *irrespective* of the perceived quality of the wine, -as say, Poor/Unacceptable; Fair/Mediocre; Good/Above Average; or Excellent to Superior-then a ranking alone might suffice. This said, the distinct limitation of using ranks

instead of scores is the fact that very different wine scores can and will receive the same relative rank ordering.

In order to understand better how ranks compare to scores, consider the ratings of 5 tasters of the same three wines as the following:

	WINE TASTER:				
WINE:	One	Two	Three	Four	Five
A	60	70	90	94	96
B	62	73	92	96	98
C	66	76	96	98	100

Since each judge has placed her/his three wines in the exact same order, each taster receives identical ranks of 1, 2, and 3, respectively. This produces 100% agreement among the four tasters. In distinct contrast, when the overall reliability of the wine scores is calculated, it produces a near-zero value of only 0.03, that is of no statistical or practical significance whatsoever. As one can detect by eyeballing the data (the so-called "intra-ocular traumatic test" or "the eyes have it test") it is clear that there is a real dichotomy in the data, with judges 3, 4, and 5 in very close agreement with each other; but, in substantial disagreement, i.e., poor or mediocre agreement, with the remaining two judges.

It needs to be emphasized that when one is interested in the perceived quality of a given wine-the usual desideratum- then the only choice is to use wine rating scores. The application of ranking methods, by this reasoning, would seriously compromise the accuracy of results reported on: the famed 1976 Paris blind tastings of Cabernet/Bordeaux and Chardonnay (Hulkower, 2009); the follow-up to the Paris tasting three decades later (Robinson, 2006); as well as the reported results of this South African tasting, by Hulkower, (2013). It is important to note here that one actually gets both a quality rating, *as well as,* a ranking, by using scores. As an example, suppose a given expert rates four wines as 95, 85, 75, and 65. These scores can be interpreted, as follows: The wine ranked number one received a score of Superior; followed by wines of Good, Fair, and Poor

quality, with overall respective rank orderings of first, second, third, and last, respectively.

A major question is where does one proceed from here? A major issue that needs to be answered in follow-up research is whether the levels of inter-rater reliability reflect very low overall and specific inter-rater levels of agreement , or whether, as was shown for the famed 1976 Paris taste-off, there was a subset of raters that were highly consistent wine judges for both white and red wines (Cicchetti, 2006).

As a final point, and consistent with wine folklore, the evaluation of wines-the very scores we assign to them-reflects a judgment that is very subjective and very personal. It's not so much what a wine tastes like as much as it is what we, in fact, prefer to drink.

So I end this with a tip of my enological hat to the wine producers, wholesalers, retailers, and imbibers who make the iffy-shifty-ephemeral phenomenon known, more informally, as the enjoyment of our next glass of wine, possible.

References

Bartko, J.J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin,* 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.

Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics,* 31, 651-659.

.Fleiss, J.L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of agreement. *Educational and Psychological Measurement,* 33, 613-619.

Fleiss, J.L., Levin, & Cho Paik, M. (2003). *Statistical Methods for Rates and Proportions*New York, NY: Wiley . (3rd ed.)

Hulkower, N.D. (2009). The judgment of Paris according to Borda. *Journal of Wine Research*, 20, 171-182.

Hulkower, N.D. The Stellenbosch tasting according to Borda. Paper presented at the Seventh Annual Conference of the American Association of Wine Economists (AAWE), Stellenbosch, South Africa, June, 2013.

Robinson, J. (2006). California triumphs again at Judgment of Paris re-run. *Purple Pages*, 1-2.

.Shrout, P.E., & Fleiss, J.L (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin,* 22, 420-428.

Table 1

Eight Sauvignon Blanc South African Wines at the 2013 AAWE Blind Tastings:

	Wine							
	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**
Cellar	De M	Du T	Lut	FS de F	Du T	GP	BR	Du T
District*	Stell	Wor	Olif	Fran	Wor	Dar	Paarl	Wor
Year	2012	2012	2010	2011	2012	2012	2011	2012
Alcohol %	13.5	12.5	12.5	13.5	12.5	13.5	13.5	12.5
U. S. Price ($)	7.00	3.50	3.50	4.50	3.50	6.80	10.00	3.50

*De M=De Morgenzon; Du T=Du Toitskloof Cellar; Lut=Lutzville; GP=Groote Post; and BR=Boschendal Reserve.

**Stell=Stellenbosch; Wor=Worcester; Olif=Olifantsriver; Fran=Franschhoek; Dar=Darling, and Paarl=Paarl.

Table 2

Eight Pinotage South African Wines at the 2013 AAWE Blind Tastings:

	Wine							
	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**
Cellar*	Anur	Du T	Stey	Clos	Du T	Aren	Du T	KWV
District**	Paarl	Wor	Stell	Stell	Wor	Rob	Wor	Paarl
Year	2008	2011	2008	2008	2011	2008	2011	2012
Alcohol %	15.0	14.5	14.5	14.0	14.5	14.0	14.5	13.5
U. S. Price ($)	7.00	3.80	22.40	15.00	3.80	8.00	3.80	5.20

*Anur= Anura Reserve; Du T=Du Toitskloof; Stey=Steytler;Clos=Clos Malverne Reserve; Aren=Arendskloof; and KWV=KWV-Café` Culture BIB.

** Paarl=Paarl; Wor=Worcester; Stell=Stellenbosch; and Rob=Robertson.

Table 3

Results of Sauvignon Blind Wine Tastings: Stellenbosch 2013 AAWE Meetings

Taster Number	W01	W02	W03	W04	W05	W06	W07	W08
1	80	80	84	80	84	89	82	80
2	78	70	85	72	75	75	85	80
3	70	78	75	80	82	78	88	85
4	70	65	83	77	86	80	75	72
5	75	60	61	88	78	69	65	89
6	73	80	80	70	72	79	80	79
7	81	83	81	82	85	84	85	84
8	88	75	90	76	88	79	80	87
9	85	80	80	90	90	85	95	90
10	95	97	88	85	90	91	92	97
11	82	80	88	81	85	87	89	91
12	89	58	65	50	75	84	50	55
13	83	81	75	80	78	77	70	82
14	73	83	70	80	78	82	70	78
15	72	83	79	81	79	75	81	86
Totals:	1194	1153	1184	1172	1225	1214	1187	1235
Averages:	79.6	76.9	78.9	78.1	81.7	80.9	79.1	82.3

Table 4

Results of Pinotage Blind Wine Tastings: Stellenbosch 2013 AAWE Meetings

Taster Number	R01	R02	R03	R04	R05	R06	R07	R08
16	90	85	80	90	70	72	83	85
17	60	80	75	60	82	60	76	78
18	70	85	82	80	85	79	80	75
19	82	80	89	65	70	65	80	85
20	90	70	90	70	80	80	85	80
21	60	55	65	73	70	65	68	60
22	79	89	92	88	90	87	79.5	82
23	75	77	68	82	80	65	70	74
24	85	65	60	70	70	72	70	75
25	82	70	72	75	85	65	60	95
26	60	70	60	60	65	60	65	70
27	81	75	74	70	75	72	89	75
28	81	74	88	85	80	68	77	90
29	85	79	75	70	85	87	89	75
30	85	75	81	82	83	85	75	74
Totals:	1165	1129	1151	1120	1170	1082	1146.5	1173
Averages:	77.7	75.3	76.7	74.7	78.0	72.1	76.4	78.2