Journal of Wine Economics, Volume 4, Issue 2, Winter 2009, Pages 233–241

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

How Expert are "Expert" Wine Judges?*

Robert T. Hodgson[a]

Abstract

Recent papers by Hodgson (2008) and Gawel and Godden (2008) have questioned the consistency of expert wine judges in a wine competition setting. In the latter paper, a methodology introduced in psychometric research to measure judge reliability corrected for chance was used to quantify judge consistency (Cohen, 1968). This paper extends that notion, suggesting a value of 0.7 for Cohen's weighted kappa might be used to define an expert wine judge. With that criterion, less than 30% of judges who participated in either of the two studies would be considered "expert." (JEL Classification: C1, L15)

I. Introduction

After the 1976 Paris wine tasting that compared French Bordeaux to California Cabernets as well as French versus California Chardonnays, several investigators questioned the reliability of the judges as well as their conclusions by employing an array of statistical methods (Ashenfelter and Quandt, 1998; Cicchetti, 2004). As a generalization, reliability was equated with concordance, meaning that reliable judges agreed with each other. Using this paradigm, Cicchetti (2004) examined the results of the 1976 Paris tasting and was able to separate more reliable tasters from less reliable ones.

Earlier, Cohen (1960) described a situation, "which arises in the clinical-social-personality areas of psychology," where judges diagnose the symptoms of an array of patients. Cohen set before him the task of evaluating the reliability of judges, meaning, to what extent would two or more judges render a concordant diagnosis allowing for the possibility that agreement in diagnoses between judges could occur by chance alone.

This paper extends Cohen's ideas to evaluate the consistency of a single judge, as contrasted to the concordance between two or more judges. To do this we ask, "How often does a judge render an identical verdict for an identical wine, and how often would identical

* I would like to thank Karl Storchmann (the editor) and an anonymous referee for their critical suggestions on this manuscript.

[a] Professor emeritus, Humboldt State University, Arcata, California. Owner/winemaker, Fieldbrook Winery. Email: bob@fieldbrookwinery.com.

© The American Association of Wine Economists, 2009

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

verdicts be expected by chance alone?" We use the results of the 2009 California State Fair commercial wine competition, in which expert wine judges were served replicate samples of wine poured from the same bottle as discussed by Hodgson (2008).

II. Methods

A. *Data Collection at the California State Fair*

The California State Fair commercial wine competition typically evaluates about 3000 wines per year using 16 panels of four judges. Each panel evaluates approximately 30 wines per flight. Within these 30 wines, several replicates exist for the purpose of evaluating each judge's consistency. For example, within a flight of 30 wines, all the judges on a panel might have samples coded #3157, #4128 and #3945—poured from the same bottle.

The judges rate the wines on a 10 point ordinal scale ranging from *No Award* to *Gold-Plus*: X, B–, B, B+, S–, S, S+, G–, G, G+. The symbols X, B, S, and G stand for *No Award*, *Bronze*, *Silver* and *Gold*.

Each judge is presented with four triplicates that are coded as Z_{jkm} where j = 1, 2, 3, 4 is the index referring to each judge on a panel, k is the index referring to each of the four wines and m(1, 2, 3) refers to each sample of the kth wine. Thus, referring to Table 1, the

Table 1
An Example Matrix of Four Judges Evaluating Four Triplicate Wines

		j = 1	*j = 2*	*j = 3*	*j = 4*
k	m				
1	1	B	G	X	X
1	2	S	B	X	G–
1	3	B	S	S–	S–
2	1	B	B	X	X
2	2	B	X	X	X
2	3	X	B	X	X
3	1	B	B	S	S+
3	2	B	S	X	S–
3	3	B	X	S–	X
4	1	X	X	X	X
4	2	X	X	X	X
4	3	X	X	X	X

An example of the Z(j,k,m) scores awarded by a panel at the California State Fair commercial wine competition. Each column represents a single judge. Each of the four test wines is identified by the index k. The triplicate samples are associated with the index m. The third judge (j = 3) scored the three samples of the first wine (k = 1) with two no-awards and one silver-minus (X, X, S–). Judge 2 scored the third wine (B, S, X), etc. All 12 samples of each wine were poured from the same bottle.

scores given by the first judge, Z_{1km}, would be (B, S, B), (B, B, X), (B, B, B) and (X, X, X). The scores of the second judge (Z_{2km}) on that panel are (G, B, S), (B, X, B), (B, S, X) and (X, X, X). For every test wine, all three replicates were served on the same flight.

B. *Calculating the Expected and Observed Frequencies for a Single Judge*

As noted above, the original data consisted of 10 ordinal scale categories, *No Award* to *Gold-plus*. Based on preliminary calculations, which revealed little justification for 10 categories as well as for computational efficiency, we reduced the number of categories to just four, *No Award* (X), *Bronze* (B), *Silver* (S) and *Gold* (G). Thus, the original categories of B–, B and B+ were aggregated to B, etc. From the four wines served in triplicate samples we compiled three data matrices for each judge. For each triplicate, we compared the first score with the second score ($Z_{jk1}:Z_{jk2}$), the first score with the third score ($Z_{jk1}:Z_{jk3}$) and the second score with the third score ($Z_{jk2}:Z_{jk3}$). Thus, for J1, the first judge (j = 1), we have the three matrices shown in Table 2 that allow for 12 comparisons.[1]

Table 3, the Observed Frequency Matrix for J1 can be deduced directly from Table 2 and consists of one occurrence of *Silver-Bronze* (S-B), one occurrence of B-S, five occurrences of B-B, three occurrences of X-X and two occurrences of B-X.

Whereas Cohen compared the decisions of one individual to another (or peer group), we compare the decisions of a single individual's scores of identical wines against that same individual's tendency to distribute awards in a given ratio. Looking at the trace elements in Table 3, we see that J1 gave the same wine an identical score eight times out of 12: three times for X : X and five times for B : B. Could this be explained by chance, or did J1 correctly identify identical wines by skill?

Table 2

Twelve Pairs of Scores for the Triplicate Samples Evaluated by J1 in Table 1

Z_{111} : B	Z_{112} : S	Z_{111} : B	Z_{113} : B	Z_{112} : S	Z_{113} : B
Z_{121} : B	Z_{122} : B	Z_{121} : B	Z_{123} : X	Z_{122} : B	Z_{123} : X
Z_{131} : B	Z_{132} : B	Z_{131} : B	Z_{133} : B	Z_{132} : B	Z_{133} : B
Z_{141} : X	Z_{142} : X	Z_{141} : X	Z_{143} : X	Z_{142} : X	Z_{143} : X

For each judge, the three scores (m = 1, 2, 3) of every kth test wine allow for three comparisons: the first score compared to the second, the first score compared to the third and the second score compared to the third. Thus for the four test wines, there are 12 pairs of scores as indicated above. The three samples of each test wine were poured from the same bottle. The original scores in Table 1 have been aggregated to X, B, S and G.

[1] For each of the three triplicate scores, there are only two independent differences, a point raised by one of the reviewers. Thus for the 12 comparisons, only eight are independent. While this might introduce a bias into the estimate of Cohen's kappa, the author believes it to be negligible in terms of practical significance.

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

Table 3
Observed Frequency Matrix for J1

	X	*B*	*S*	*G*
X	3	0	0	0
B	2	5	1	0
S	0	1	0	0
G	0	0	0	0

The observed frequency matrix for J1 is computed directly from Table 2. There were two occurrences of B : X, five occurrences of B : B etc. This judge gave the same wine identical scores 8 times out of 12.

To answer this we need to know in what ratio does J1 ordinarily give *No Awards*, *Bronzes*, *Silvers* and *Golds*. To assess these ratios, we look at a J1's scores from contiguous flights of the same varietal. For example, the data in Table 3 were taken from the second flight of Cabernet Sauvignons. Table 4 lists the distribution of awards given by J1 (as well as the other judges on the panel) for all 85 Cabernet Sauvignons served on flights 1, 2, and 3 of the same day. Note that the medal distributions of the four judges are significantly different. Thus J1 typically awarded *No Awards* 48 percent of the time (41/85), *Bronze* 22 percent of the time, etc.

Based on these 85 observations, one can calculate the expected frequency matrix for Judge 1 for the 12 test cases. Let f_i, f_j $(i,j = \mathrm{X, B, S, G})$ denote the observed frequency distribution of awards given by the values listed in Table 4. Thus, $f_X = 41$, $f_B = 19$, $f_S = 16$ and $f_G = 9$. Let f_T be their sum 85. Then the expected frequencies for Judge 1 are given by

$$f_{ij} = f_i \cdot f_j \cdot N/f_T^2 \tag{1}$$

where $N = 12$. For example,

$$f_{xs} = f_x \cdot f_s \cdot N/f_T^2 = 41 \cdot 16 \cdot 12/85^2 = 1.09$$

The completed expected frequency matrix is shown as Table 5.

Table 4
Frequency of Medals Awarded by the Four Judges

	J1	*J2*	*J3*	*J4*
None	41	44	57	57
Bronze	19	15	6	4
Silver	16	17	16	18
Gold	9	9	6	6

In contiguous flights of Cabernet Sauvignon, J1 awarded 41 No Awards, 19 Bronze Awards, 16 Silver Awards and 9 Gold Awards. Note the ratio of awards given by J1 are similar to those of J2, but quite different than J3 or J4.

Table 5
Expected Frequency Matrix for Judge 1
(n = 12)

	X	*B*	*S*	*G*
X	2.79	1.29	1.09	0.61
B	1.29	0.60	0.50	0.28
S	1.09	0.50	0.43	0.24
G	0.61	0.28	0.24	0.13

For the 12 replicate samples, the expected frequency for J1 to award B:X is 1.29, the expected frequency to award the same wine B:B is 0.60 as calculated from equation 1. The sum of the trace elements (3.95) is the expected frequency of J1 awarding identical scores for 12 pairs of samples based on chance alone.

C. *Cohen's Kappa*

In the foregoing example, Cohen's kappa statistic is used to see if the eight correct identifications of an identical wine could have occurred by chance alone. Cohen's kappa is obtained using a modified chi-square statistic,

$$\kappa = \frac{(f_o - f_c)}{(N - f_c)} \tag{2}$$

where f_o is the sum of the trace elements of the observed frequency matrix, f_c is the sum of the trace elements of the expected frequency matrix and N is the number of observations (12). As Cohen (1960) points out, this form of the statistic emphasizes agreement instead of disagreement.

In a later paper, Cohen (1968) realized that one might give partial credit for "near misses." For ordinal scale data, a *Bronze-Silver* mismatch in scores is not as bad as a *No Award–Gold*. Thus, Cohen introduced weighted kappa, κ_w, to atone for less-than-perfect agreement. The computation of weighted kappa uses a weighting matrix to include the off-diagonal elements of the observed data matrix (Table 3), whereas "straight" kappa does not. The maximum value of kappa (or weighted kappa) is unity, corresponding to a judge who is perfectly consistent whereas judges whose scores could be explained by chance alone would have corresponding values of kappa less than or equal to 0.

Using the same weighting matrix (Table 6) as Gawel and Godden (2008), Table 7 lists the values of kappa and weighted kappa for the data in Table 1. For J1 there was a 0.5 (50%) chance that the eight correct "match-ups" of the identical wines could have occurred by chance. Conversely, there was a 50% chance that the correct match-ups did not occur by chance, but were based on skill. The value of weighted kappa for J1 is 0.81, which gives

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

Table 6
Weighted Kappa Matrix

	N	*B*	*S*	*G*
N	1	.8	.2	0
B	.8	1	.8	.2
S	.2	.8	1	.8
G	0	.2	.8	1

Table 7
Kappa and Weighted Kappa for the Data in Table 1

	J1	*J2*	*J3*	*J4*
Kappa	0.50	−0.03	0.34	0.16
Weighted	0.81	0.36	0.20	0.14

credit for near misses. This suggests there was an over 80 percent chance that the match-ups (or near match-ups) were based on skill, not chance.

III. Results

Figures 1 and 2 display the relative cumulative frequency plots for the 2009 California State Fair (CSF) commercial wine competition as well as the results of Gawel and Godden (2008), which used Cohen's weighted kappa to evaluate the consistency of Australian wine judges. As Gawel and Godden found significant differences between white and red Australian wines, we have also separated our results for easy comparison. Figure 1 compares the results for white wines, Figure 2 for reds.

The California judges represent a broad cross section of evaluators: faculty of enology and viticulture, winemakers, wine buyers for major retailers and wine journalists. All have participated in a wine evaluation course offered by the University of California and have passed a qualifying examination.

The Australian data were extracted from Figure 3 of Gawel and Godden (2008). Their observations cover 571 judges who had undergone a training program administered by the Australian Wine Research Institute. The study covered a 15-year time period where between 20 and 40 wines were served to each judge between two and three days apart in flights of 14 to 37.

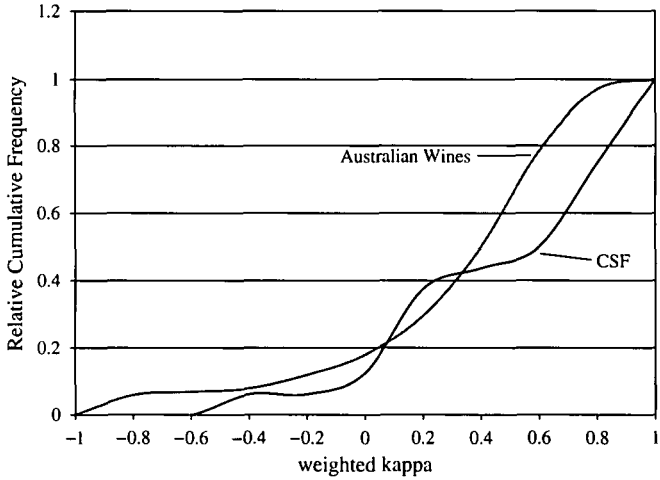
While it was not the purpose of this study to contrast the Australian and American judges, it is interesting to compare the two studies. Whether reds or whites, Australians or Americans, less than 30% of the judges are "expert" using a weighted kappa of 0.7 as a criterion.

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

Figure 1

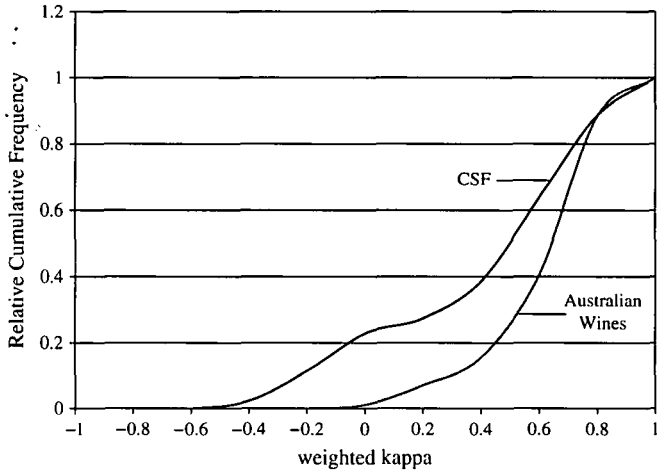
Relative Cumulative Frequency vs. Weighted Kappa White Wines



Australian data extracted from Gawel and Godden (2008). California State Fair (CSF) data from the 2009 California State Fair commercial wine competition. Approximately 25% of the California judges and 10% of the Australian judges would be considered "expert" using a criterion of $\kappa_w \geq 0.7$.

Figure 2

Relative Cumulative Frequency vs. Weighted Kappa Red Wines



Australian data extracted from Gawel and Godden (2008). California State Fair (CSF) data from the 2009 California State Fair commercial wine competition. Approximately 10% of the California judges and 30% of the Australian judges would be considered "expert" using a criterion of $\kappa_w \geq 0.7$.

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press

IV. Discussion

To what standards should we hold expert wine judges? Cohen's kappa supplies an excellent tool to evaluate judge consistency. The direct interpretation is that a value of zero or less means that the decision of the judge was likely based on chance alone; for a kappa of unity, the judge would be considered excellent, being perfectly consistent. Exactly what criterion might separate expert judges from others is not clear, but I suggest that a value of weighted kappa between 0.5 and 0.7, discussed below, as a possibility.

It is easy to simulate a judge's level of expertise under the conditions of evaluating four triplicate samples. Starting with perfect consistency, just how far from perfection might a judge veer in order to maintain a weighted kappa above 0.7, or above 0.5?

The value of κ_w decreases rapidly depending on the magnitude of the miscategorization as well as the number of times it occurs. Suppose a judge is perfectly consistent in three of the replicate wines, but misclassifies one triplicate sample by a single medal, e.g., BBB, SSS, BSB and SSS. We can label this sequence as specifying the *within-group-difference* as *wgd*(0, 0, 1, 0), the "1" identifying the single medal difference (bronze-silver) for the third triplicate. As another example, we could code BGS, GGG, SBS and GGG by *wgd*(2, 0, 1, 0).

Weighted kappa depends on the errors in classification, the expected frequencies and the weighting matrix. For the weighting matrix in Table 6 and using expected frequencies of 5%, 10%, 25%, 60% for G, S, B, X, $\kappa_w(0, 0, 1, 0) = 0.87$. Continuing, we find $\kappa_w(0, 0, 1, 1) = 0.73$, $\kappa_w(0, 1, 1, 1) = 0.60$ and $\kappa_w(1, 1, 1, 1) = 0.46$. The case where a judge is consistent for three triplicates, but miscategorizes one triplicate by two medals, say BGB or NSS, we find that $\kappa_w(0, 0, 0, 2) = 0.46$. All other medal combinations yield values of kappa less than 0.46. These results are summarized in Table 8.

If using weighted kappa as a criterion, what might characterize an expert judge? To maintain a weighted kappa score of 0.7 or above, an *expert* judge must be perfectly consistent with the exception of two misclassifications of a single medal group, for example: BBB, NNN, BSS, SGG. Any two-medal misclassification, say BGG, would exclude a judge from this "expert" classification. By lowering the weighted kappa criterion to 0.5, an "expert" judge could misclassify four wines by a single medal group or a single wine by two categories, but no more. For example, the case of BBG, SSB, SSS, NNN (One two-medal misclassification and one single medal misclassification) would exclude a judge from expert status.

In addition to being consistent, an expert judge must be able to discriminate wines over a broad range of quality. (It is easy to achieve a kappa of unity by scoring all wines identically). To measure discrimination, one might impose, *a priori*, a medal distribution profile. For example: 60% no awards, 25% Bronze, 10% Silver, and 5% Gold. Another would be to use the observed distribution of medals for the entire competition as a basis for a measure of concordance. This hints at a two parameter weighted kappa, $\kappa_w(r, s)$ where r would indicate consistency and s would refer to concordance, i.e., the normal interpretation of kappa.

Table 8
Weighted Kappa as a Function of Within-Group-Difference and Frequency of Occurrence

	within group difference			
N	0	1	2	3
0	1	1	1	1
1	1	0.9	0.5	0.3
2	1	0.7	−0.1	−0.3
3	1	0.6	−0.6	−1.0
4	1	0.5	−1.0	−1.7

Within group difference (*wgd*) refers to the difference in nominal scale rankings. A *wgd* of 1 corresponds to X-B, B-S, or S-G. A *wgd* of 2 corresponds to X-S or B-G. A *wgd* of 3 corresponds to X-G. The column under *N* refers to the number of times the *wgd* occurs in four replicated samples. For example if a judge scored four triplicate wines as (GSG, SSS, BSB, BBB), the maximum *wgd* is 1 and the frequency of occurrence is $N = 2$. This corresponds to a weighted kappa of 0.7.

V. Conclusion

What do we expect from expert wine judges? Above all, we expect consistency, for if a judge cannot closely replicate a decision for an identical wine served under identical circumstances, of what value is his/her recommendation? In addition we expect an expert wine judge to have the ability to discriminate across a broad range of quality. Based on a weighted kappa criterion of 0.7, this study as well as that of Gawel and Godden (2008) suggests that less than 30% of *expert* wine judges studied are, in fact, "expert."

References

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin*, 70(4), 213–220.

Cicchetti, D.V. (2004). Who won the 1976 blind tasting of French Bordeaux and U.S. Cabernets? Parametrics to the rescue. *Journal of Wine Research*, 15(3), 211–220.

Gawel, R. and Godden, P.W. (2008). Evaluation of the consistency of wine quality assessments from expert wine tasters. *Australian Journal of Grape and Wine Research*, 14, 1–8.

Hodgson, R.T. (2008). Evaluation of wine judges at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.

https://doi.org/10.1017/S1931436100000821 Published online by Cambridge University Press