EVALUATION OF THE CONSISTENCY OF WINE QUALITY ASSESSMENTS FROM EXPERT WINE TASTERS

RICHARD GAWEL[1,2,3] and PETER W. GODDEN[1]

1 The Australian Wine Research Institute, PO Box 197, Glen Osmond, 5064, Australia
2 Cooperative Research Centre for Viticulture, PO Box 154, Glen Osmond, 5064, Australia
3 Corresponding Author: Mr Richard Gawel, facsimile: +61 8 83036601,email: richard.gawel@awri.com.au

Abstract

The within-assessor consistency of 571 experienced wine tasters in assigning quality ratings to red and white table wines was determined. Wine quality scores were collected over a 15 year period from tasters undertaking an advanced training course in wine quality assessment. Consistency was measured by correlating the scores given to duplicate presentations of wines, calculating the pooled variation in repeat scores, and by assessing the tasters' ability to allocate duplicate presentations of the same wine to the same quality category. The distribution of individual assessor correlation coefficients for both red and white wines was left skewed with a median of 0.48 and 0.37 respectively. The ability of the tasters to consistently allocate red wines to the same or similar quality categories was particularly good. Consistency was improved by combining the independently assigned scores of three assessors as is done in the Australian wine show system. Assessors generally showed greater reproducibility in scoring red wines compared with whites, and in general, the ability of a taster to consistently score red wines was a poor predictor of their ability to consistently score white wines, and vice versa. Lastly, while the majority of wine tasters showed statistically significant scoring consistency, there was considerable variation between individuals in their ability to do so.

Keywords: wine quality, wine scoring, wine assessment, taster consistency

Introduction

Quality in foods and beverages including wine can been defined as "the ability of a set of inherent characteristics of a product, system or process to fulfill requirements of customers and other interested parties" (ISO 2000). The task of making purchase decisions based on wine quality considerations can be challenging for many consumers. This is probably due to the large number of viticultural and winemaking processes which impact on wine style and character, and on the variety of contexts in which a wine may be consumed. For this reason, many wine consumers rely on expert opinion to guide them in their purchase choice. Expert guidance can take the form of an individual opinion, either as a description, quality score (usually out of a maximum of 20 or 100), or quality range usually designated by 'star' ratings. Some consumers are also guided in their purchasing decisions by wine show results which are based on the combined quality ratings of a small team of expert tasters. In Australia wine show results are usually reported to the consumer in the form of a specific type of medal i.e. gold, silver or bronze (Dunphy and Lockshin 1998).

Evaluation of wine quality is usually undertaken by 'wine experts' as their experience and training enables them to both identify wine defects and also evaluate whether the wine being assessed typifies the variety, region or style which it represents. While the experts may agree in general terms as to what sensory aspects contribute positively and negatively to overall wine quality, this does not guarantee that individual assessors will weight the different underlying

dimensions of wine quality in a similar way when arriving at an overall quality score (Cliff and King 1999). Attempts have been made to circumvent this issue by prescribing weights to the different facets of wine quality (Ough and Baker 1961, Rankine 1986). However, this approach has been criticised on the basis that the assigned weights are necessarily arbitrary (Lawless et al. 1997). While this matter remains unresolved, it could be argued that good intra-assessor consistency is a necessary prerequisite of valid assessor ratings, with inter-assessor agreement being another. That is, for a quality score to be of any value, an assessor should firstly demonstrate an ability to reproduce that quality score over repeat assessments of the same wine.

Given that expert scores are frequently used to promote the various merits of commercial wines, there are surprisingly few published studies that quantify the ability of expert tasters to consistently score wine quality (Ough and Baker 1961). This is possibly due to the fact that replicate presentations of wines are not routinely given during wine competitions or other tastings where expert tasters are involved in tasting significant numbers of wines. Brien et al. (1987) summarised the results of a number of replicate wine-tastings involving highly experienced winemakers scoring wines made using grapes grown under different viticultural treatments but vinified in an identical fashion. Although the quality of their wines would have probably only varied slightly, the judges were mostly consistent in their appraisal. Similarly, the majority of a group of Australian winemakers were able to consistently discriminate between the quality levels of a Chardonnay wine that had been aged in various types of oak barrels (Gawel et al. 2001). Lawless et al. (1997) measured the consistency of 21 expert judges in scoring 14 commercial Sauvignon Blanc wines using a structured 20 point quality scale, and reported a range of individual assessor correlations of between around 0.1 and 0.85. All these authors employed correlation coefficients to measure consistency. This is not the ideal approach as this statistic measures association rather than agreement, and is therefore unaffected by systematic scoring shifts.

In many instances quality outcomes are most appropriately communicated by a category representing a score range rather than the score itself. For example, wine competitions award medals to wines based on the range in which their quality score falls (See Table 1 for an example of the Australian wine show system). In these circumstances, using correlation coefficients to measure consistency is not ideal as the statistic is adversely affected by variations in repeat scores given within the same medal range. It could be argued that from the stand-point of wine show judging, that providing scores to a wine on repeat tastings that both fall within the same medal range is in fact the optimal result. Therefore, alternate approaches to measuring consistency such as the weighted Cohen's Kappa (κ) statistic are more appropriate in these circumstances as κ measures agreement of ordinal data while also accounting for the size of disagreement. Most importantly, κ is a true measure of consistency as it reduces when an assessor systematically shifts their scores on one tasting occasion relative to another. On the other hand, a high κ statistic can be obtained by scoring all the wines in a narrow range regardless of distinctions in quality.

In this longitudinal study wine quality score data was collected over a 15 year period from 571 experienced wine assessors. Their consistency in scoring both dry red and dry white commercial table wines was measured using Pearson's correlation coefficient, 2) Cohen's weighted κ and 3) a simple mean of the absolute difference between repeat scores. The tasters' ability to discriminate between wines on the basis of perceived quality was also assessed. The score-rescore consistency of panels of three judges as commonly used in the Australian wine show system was also determined and compared with individual performance.

Methods

The assessors

All assessors were participants of a four day advanced wine assessment course conducted by the Australian Wine Research Institute. The course was in part developed to further train experienced wine tasters working within the Australian wine industry in the

skills associated with formal show judging. A total of 571 assessors took part over a 15-year period, with between 29 and 32 taking part at any one time. Demographic information was obtained from the last 120 participants. 75% of these were practicing winemakers, 14% were from the commercial wine trade, 8% were wine researchers and 3% were wine journalists. All would be considered 'wine experts' using the criteria of Parr et al. (2003). Furthermore, a random selection of 50 participants from the list of all participants over the 15 year period revealed that all were expert wine tasters under the criteria of Parr et al. (2003). This was not unexpected given that the course was marketed exclusively to winemakers and other wine professionals and that a high degree of wine tasting experience was an essential course prerequisite.

The wines used to assess consistency

All individuals from within each group of assessors were provided with the same duplicate presentations of between 14 and 34 commercial Australian red wines, and between 14 and 37 Australian dry white wines (means of both red and white wines = 23). The test wines for each course were selected by the course organisers using the criteria that 1) they represented a diverse range of grape varieties and styles (light, medium and full bodied), and 2) the selected styles and varieties were likely to be familiar to those undertaking the course. The presented varieties remained largely unchanged across the 15 year period with white wines being heavily represented by Chardonnay, Semillon, Sauvignon Blanc and Riesling and their blends, and red wines by Shiraz, Cabernet Sauvignon, Grenache, Pinot Noir and Merlot and their blends.

The wines and their duplicates were presented over a three or four day period with each duplicate typically being presented two or three days apart. The test wines were embedded within large flights of wines representing a particular variety or style. Tasters were told of the variety and style of the different flights of wines, and had previously discussed the sensory attributes which defined good examples of that style or variety. Although the tasters were aware that their judging performance was being assessed, they were unaware of the identity of the wines, or which of the presented wines were being used to assess their consistency.

Tasting conditions and scoring system

All assessments were conducted in white booths under fluorescent lighting. Communication between assessors while scoring was prohibited. Approximately 30-40 mls of wine were presented in the same order at room temperature in clear ISO tasting wine glasses. The assessors scored the wines for overall quality using a twenty point scoring system incremented in half points. In arriving at a quality score, tasters were allowed to weight what they considered to be the various aspects of wine quality relating to that style or variety in any way they saw fit. The scale was broadly structured with details given in Table 1. In general, the scoring ranges were equivalent to those used in the Australian wine show system. If any wine was perceived to be affected by cork taint or random oxidation, assessors were asked to indicate this on their score card and not score the wine.

Statistical analysis

Assessor consistency: Consistency between repeat evaluations of the same wine by an individual were estimated using 1) Pearson's correlation coefficient, which will hereafter be referred to as "reliability" after Brien et al. (1987), 2) the mean of the absolute difference between scores allocated to the same wine (average absolute difference, AAD) and 3) Cohen's weighted κ. The latter was weighted such that allocating wines to the same or adjacent scoring range over repeat evaluations was heavily weighted while all other categorisations were very lightly weighted (Table 2).

Assessor discrimination: As high consistency can be achieved by using a narrow range of the scorecard it is necessary to assess the tasters ability to separate the wines on the basis of quality. This was done by calculating the ratio of the variation amongst mean wine scores to that of the pooled variation in scores given to the same wine (Brien et al. 1987, Schlich 1994).

Small panel consistency: Nine or ten random subsets of 3 assessors were created from the 29-32 assessors undertaking each tasting course. For each tasting occasion, the independently derived scores from three assessors were summed for each wine. An overall quality category was then assigned using the criteria given in Table 1. This methodology was chosen as it reflects that of the conduct of the major Australian wine shows whereby judges independently assess wines using a 20 point scoring system and medals are awarded based on the ranges of the summed total of scores. Intra-panel consistency was determined by calculating a weighted κ statistic using the weights given in Table 2.

Consistency differences between red and white wine assessment: Differences between red and white wine reliabilities were determined by converting them to a normally distributed variable Fisher's z' transformation and calculating 95% confidence limits based on the known standard error of $(1 / \sqrt{(N-3)})$, where N = number of wines. As the number of wines used to calculate individual reliabilities was not the same across groups, the overall difference between red and white wine taster performance was determined using the distribution free sign test. The sign test was also used to assess differences between red and white wine consistency using Cohen's weighted κ. Differences in the variability of scores given to red and white wines was determined using the F test, and differences in red and white wine score distributions by the large sample Kolmogorov-Smirnov two sample test.

Changes to assessor consistency over time: This is a longitudinal study extending over a 15 year period. A runs test (Siegel 1959) was applied to determine if sufficient evidence existed for either an upward or downward trend in assessor consistency over that period.

All analyses were performed using Minitab v14.0, with the exception of Cohen's weighted κ, sign test and runs test which were done using Microsoft Excel routines. A significance level of 5% was used.

Results and Discussion

Individual Assessor Consistency

The distribution of reliabilities for red and white wine is given in Figure 1 and Table 3. The distribution was skewed to the left which is typical for distributions of correlation coefficients. Around 2/3 of assessors showed statistically significant reliabilities for red wines, while only around ½ of assessors showed significant reliabilities for white wines (Table 3).

The AAD is the difference between scores given to the same wine, averaged across all wines tasted by that assessor, and as such is a simple measure of consistency in an absolute sense. Judge reliability does not accommodate the situation whereby a judge systematically shifts their scoring across tasting occasions. That is, reliability is a measure of association rather than a true measure of consistency. A high proportion of assessors with an AAD less than 1.5 was observed (>85%) for both red and white wines (Figure 2). In the Australian wine show system, the scoring range for individual medals spans 1.5 points which suggests that most of the judges were generally consistent, at least within the context of the scoring ranges used in this system. The high proportion of assessors with a low AAD suggests a higher degree of taster consistency than that suggested by the reliabilities. This could be due to two factors. Firstly, the reliabilities are strongly affected by outliers. Correctly repeated scores at the extreme ends of the scoring scale will result in high reliability even if all the intermediate scoring pairs are essentially random. Conversely, the reliability obtained from a large set of closely matched scores across the scoring range can be significantly reduced if a single large mismatch occurred at either end of the scoring range. Furthermore, the range of scores used by experienced wine tasters was largely confined to the range 13.0 to 19.0 (data not shown). Avoidance of the extreme ends of a scoring range are expected as tasters "reserve" the high scores to "iconic" or "ideal" wines. Tasters also avoided scores below 12.0 as Australian winemakers usually reserve these scores for very poor wines, which were not presented here. Smaller AAD's would logically be expected to occur in circumstances where a narrow scoring range is being used which may explain the high proportion of AAD values below 1.5.

The fact that correlation does not account for systematic biases, and the average absolute difference is influenced by the scoring range used means that, by themselves, they have limited utility as a measure of judge consistency. On the other hand, Cohen (1960) proposed a statistic that is a measure of agreement that accounts for the percentage of times two scores for the same wine fall within the same quality range adjusted for chance. Later, a generalised version now known as Cohen's weighted κ (Cohen 1968) was introduced to accommodate the size of misclassification by incorporating weights which introduce a higher penalty for larger misclassifications. If a rater consistently allocates the same wine to identical quality ranges then κ will be near its maximum value of +1.0. If there is no consistency other than what would be expected by chance, $\kappa \leq 0$. Figure 4a gives the distribution of weighted κ's based on the weights given in Table 2. While the exact choice of weights is subjective, the weights were chosen so as to severely penalise major misclassification compared with exact matches or minor misclassification. The distributions of Pearson's r and Cohen's weighted are not directly comparable. However they both have maximum values of 1.0 indicating perfect association and agreement respectively. For red wines, the distribution of weighted was more negatively skewed than that of the reliabilities which reflects the way in which the two statistics deal with misclassification and scoring differences respectively. The amount of major (greater than 1 medal class) misclassification across all assessors was low for both red (13%) and white (16%) wines. This resulted in high weighted κ values due to the strong weightings in favour of correct and minor misclassifications. In contrast, many of the assessor reliability values would have been severely affected by even this relatively small proportion of major scoring discrepancies.

Differences in consistency of scoring red and white wines

40 assessors had significantly different correlation coefficients between red and white wines ($P < 0.05$). Of these, over three quarters (31) were found to be more reproducible when assessing red wines. Higher red wine reliabilities were recorded by 361 of the 561 assessors, which indicates a significantly higher red wine scoring reproducibility compared with that of white wine ($z = 6.28$, $P < 0.001$). In addition the absolute average difference was significantly higher ($z = 8.00$, $P < 0.001$) and the median weighted κ for all tasters across all wines was significantly lower for white wines (0.392) compared with red wines (0.644) indicating greater variability in repeat assessments of white wines compared with red wines.

All of these results strongly suggest that the ability of assessors to reproduce quality assessments of red wines was significantly higher than that of white wines. The difference may have been due to variations in the range in qualities between the red and white wines that were presented. However, the differences between the variances in scores given to red and white wines by group showed that this was unlikely to be the cause. When analysed by group, the variance in white wine scores was significantly higher than red wine scores for four groups, red wine higher than white wine for 6 groups, with no significant difference in variance for the remaining nine groups (F test, $P < 0.05$, data not shown). Furthermore, there was no significant difference in the score distributions between white and red wines ($P > 0.05$). Therefore it seems that the higher consistency of tasters in scoring red wines cannot be attributed to systematic quality variations resulting from the choice of wines that were presented. One possibility is that consistency may have been enhanced by assessors using perceived colour density as a defacto measure of red wine quality. The perceived colour of wine is also known to influence the perception of its aroma profile (Parr et al. 2003) and therefore presumably the interpretation of its overall quality. For Shiraz and Cabernet Sauvignon, which are two of the red varieties heavily represented in the courses, colour intensity has previously been shown to correlate well with the positive attributes of perceived flavour and astringency (Francis et al. 1999, Gawel et al. 2001). On the other hand, tasters cannot reliably draw conclusions as to the overall quality of a white wine from its colour except in the relatively uncommon circumstance where excessive development or oxidation is present as indicated by deep yellow or brown hues in the wine.

Only a weak association between the reliability of judges when assessing red wines compared to white wines was evident (r = 0.232). The association between weighted κ for red and white wines is similarly weak (r = 0.266) which suggests that the assessors ability to consistently score white wines for quality is a poor indicator of their ability to score red wines. The reasons for this are unclear. However, many of the assessors were practicing winemakers, so perhaps the differences in the ability of some assessors to consistently score red compared to white wines was a reflection of their different levels of production experience with the two wine types.

Assessor discrimination: Figure 3 shows the degree of assessor discrimination for both red and white wines. On average, the between wine variation in scores was 3.67 times greater than the within wine variation for red wines, and 2.5 times greater for white wines. A statistically significant 64.2% of assessors discriminated among red wines to a greater extent compared to white wines ($P < 0.001$). The proportion of significant discriminators was very similar to that of reliable judges, which is expected as both are affected by the amount of variability given to repeat evaluations of the same wine (Brien et al. 1986).

Only 3.9% of the assessors who had a significant red wine reliability, and 6.5% who had a significant white wine reliability, did not also discriminate between wines. These results demonstrate that the vast majority of the assessors achieved a consistent scoring pattern not by scoring all the wines in a narrow range, but rather by adequately discriminating between the wines on the basis of their perceived quality. Brien et al. (1987) have suggested that simultaneously high reliabilities and discrimination statistics are typically obtained by experienced and confident wine judges.

The Consistency of Panels of Assessors

Panels of tasters rather than individuals are often employed to measure wine quality. This is true of the Australian wine shows where three judges taste the competition wines independently, and then additively combine their scores to arrive at a quality rating and overall classification (gold, silver, bronze and no medal). Figure 4b gives the distribution of weighted κ values for the repeated quality classification as determined by the combined independently derived scores of three assessors. The median panel weighted κ for red wines (0.77) was significantly higher than that for white wines (0.44) ($P < 0.001$), and both were higher than that for individuals (Table 3). The panel red wine κ distribution was slightly more skewed compared with the distribution generated by individuals which also suggests greater consistency on the part of panels in allocating red wines to quality categories. However, for white wines the difference between panel performance and individual performance was less clear. White wine assessment was once again shown to be far more variable compared with red wine assessment. This was the logical result of the greater inconsistency in the individual contributions to the panel outcomes for white wines, and reinforces the notion that while the variation in sums of scores provided by a panel are expected to be lower than the variation in individual scores, this does not necessarily guarantee greater consistency in repeat evaluations. However, the higher panel weighted κ for both red and white wines indicates that, in general, the combined results of three tasters are more consistent than those achieved by individual assessors. This result supports the use of panels of tasters rather than individuals when attempting to allocate wines into broad quality categories as is done in wine shows.

Changes to assessor consistency over time

The data for this study was collected over a period spanning a decade and a half. It is possible that over this period the general ability of wine judges may have changed - either positively due to increased knowledge of wine judging processes, or negatively due to an increase in the number of different varieties and styles that progressively became available in the Australian marketplace. A runs test was used to determine whether the median reliability statistic for each course varied in a systematic fashion over time i.e. was not random. No significant variation from randomness was observed for either wine type ($P > 0.05$). The lack of any systematic change in performance could be accounted for by the consistent use of the same group of wine varietals over the time period, and the fact that the demographics of the groups remained largely

unchanged with experienced, practicing winemakers being in the majority throughout the entire period.

Conclusion

The use of an overall score to communicate the perceived overall quality of a wine is used throughout the Australian wine industry. In particular, wine quality scoring is widely employed in the context of wine show judging. We evaluated the scoring reproducibility of 561 experienced wine tasters by correlating scores given to the same set of wines over two separate occasions, and found a large amount of variation between them with regard to their consistency. While the majority of the tasters'behaviour were reproducible both in terms of replicating quality scores and allocating wines to the same quality category when presented in duplicate, some greater consistency was achieved when combining the scores of three assessors. This result reinforces the value of using teams of tasters when evaluating wine quality. The apparent greater difficulty in consistently assessing white wine compared with red wine quality needs to be explored further to ascertain whether experience with this type of wine influences performance, or that the task of white wine assessment is just inherently more difficult.

Acknowledgements

This project was supported by Australia's grapegrowers and winemakers through their investment body, the Grape and Wine Research and Development Corporation, with matching funds from the Federal Government, and by the Cooperative Research Centre for Viticulture. We also acknowledge Peter Leske for his work in developing and conducting the early Advanced Wine Assessment Courses.

References

Brien, C.J., May, P. and Mayo, O. (1987) Analysis of judge performance in wine quality evaluations. Journal of Food Science **52**, 1273-1279.

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

Cohen, J. A. (1968) Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin **70**, 213-220.

Cliff, M.A. and King, M. (1999) Use of principal component analysis for the evaluation of judge performance at wine competitions. Journal of Wine Research **10**, 25-32.

Dunphy, R. and Lockshin, L. (1998) A contemporary perspective of the Australian wine show system as a marketing tool. Journal of Wine Research **9**, 107-130.

Francis, I. L., Cynkar, W., Kwiatkowski, M., Williams, P. J., Armstrong, H., Botting, D.G., Gawel, R. and Ryan, C. (1999) Assessing wine quality with the G-G assay. Proceedings 10th Wine Industry Technical Conference; Adelaide, Australia. (Australian Wine Industry Technical Conference Inc.: Adelaide) pp.104-108.Gawel, R., Iland, P.G., Leske, P.A. and Dunn, C.G. (2001) Compositional and sensory differences in Syrah wines following juice run-off prior to fermentation. Journal of Wine Research **12**, 5-18.

Gawel, R., Royal, A., and Leske, P.A. (2002) The effect of different oak types of the sensory properties of a Chardonnay wine. Australian and New Zealand Wine Industry Journal **17**, 10-14.International Organization for Standardization (2000) ISO 9000: 2000. Quality management systems: Fundamentals and vocabulary. p29.

Lawless, H., Yen-Fei, L. and Goldwyn, C. (1997) Evaluation of wine quality using a small-panel hedonic scaling method. Journal of Sensory Studies **12**, 317-332.

Ough, C.S. and Baker, G.A. (1961) Small panel sensory evaluations of wines by scoring. Hilgardia **28**, 587-619.

Parr, W.V., White, K.G. and Heatherbell, D.A. (2003) The nose knows: Influence of colour on perception of wine aroma. Journal of Wine Research **14**, 79-101.

Rankine, B. (1986) Roseworthy develops new wine scorecard. Australian Grapegrower and Winemaker. February, 16.

Table 1: Quality scoring ranges used to allocate quality designations

Individual Score Range	Panel[#] Score Range	Description	Medal Range*
18.5 – 20.0	55.5 – 60.0	Outstanding	Gold
17.0 – 18.4	51.0 – 55.0	Very Good	Silver
15.5 – 16.9	46.5 – 50.5	Above Average	Bronze
Less than 15.5	Less than 46.5	Average or Below Average	No Medal

summed scores of three assessors

* As awarded in the Australian wine show system

Table 2: Weights used for Cohen s Weighted □

Assessement 1	< 15.5	15.5-16.5	17.0-18.0	18.5-20.0
Assessment 2				
< 15.5	1.0	0.8	0.2	0.0
15.5-16.5	0.8	1.0	0.8	0.2
17.0-18.0	0.2	0.8	1.0	0.8
18.5-20.0	0.0	0.2	0.8	1.0

Table 3: Summary of individual assessor statistics for red and white wine sets

	Individual Assessors (n=571)					
	Reliability (Pearson's r)		Average Absolute Difference		Weighted Cohen's □	
	Red Wines	White Wines	Red Wines	White Wines	Red Wines	White Wines
Mean	+0.451[b]	+0.351[a]	1.04[a]	1.16[b]	+0.593	+0.269
Median	+0.476	+0.369	1.02	1.13	+0.644	+0.392
Minimum	-0.387	-0.416	0.14	0.49	-0.671	-1.000
Maximum	+0.967	+0.971	2.18	2.69	1.000	+1.000
% performing^ assessors	67.6	51.1	91.1	84.9	-	-

Means with different subscripts are significantly different ($P < 0.05$)

^ Performing assessor defined as one with an AAD < 1.5, or a statistically significant reliability ($P < 0.05$)

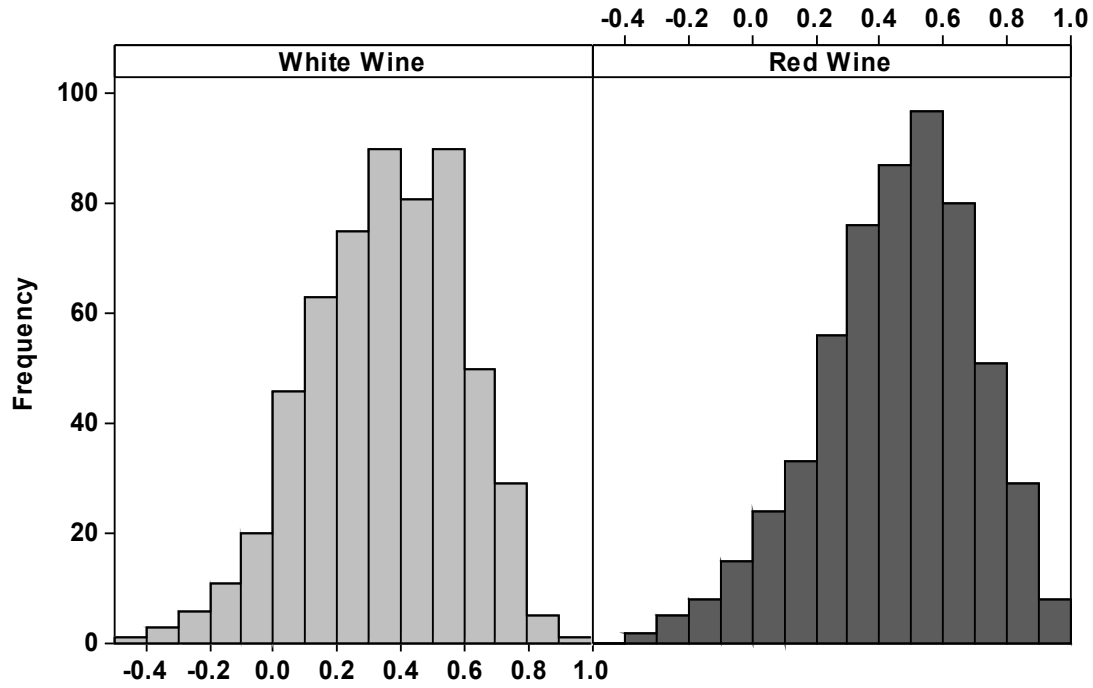

Figure 1: Distribution of reliabilities of individual assessors as measured by Pearson's correlation coefficient applied to scores given to the same wine on duplicate presentations (n = 571).

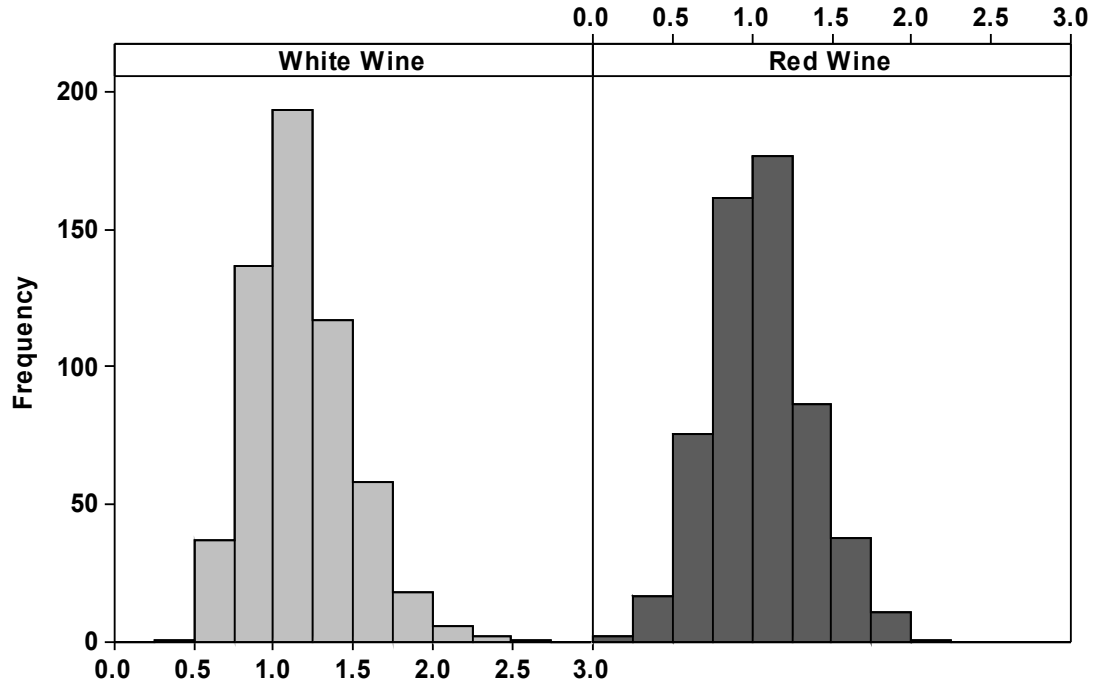

Figure 2: Distribution of the absolute average difference between scores given to the same wine on duplicate presentations (n = 571).

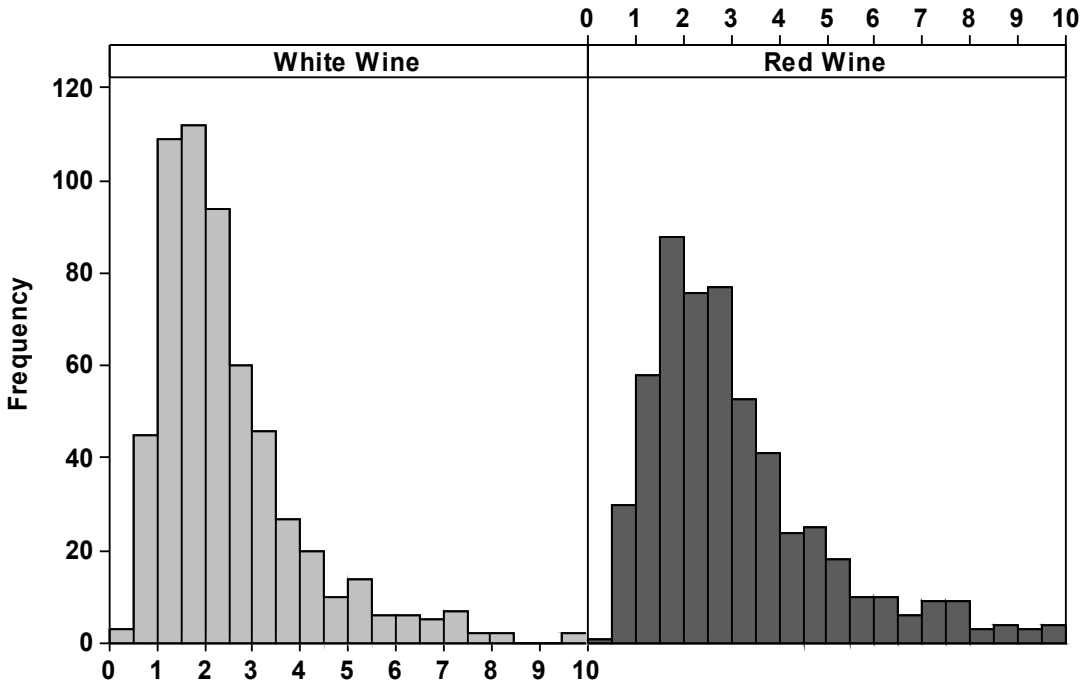

Figure 3: Distribution of the ratio of between wine to within wine variation (discrimination) by assessor (n = 571)

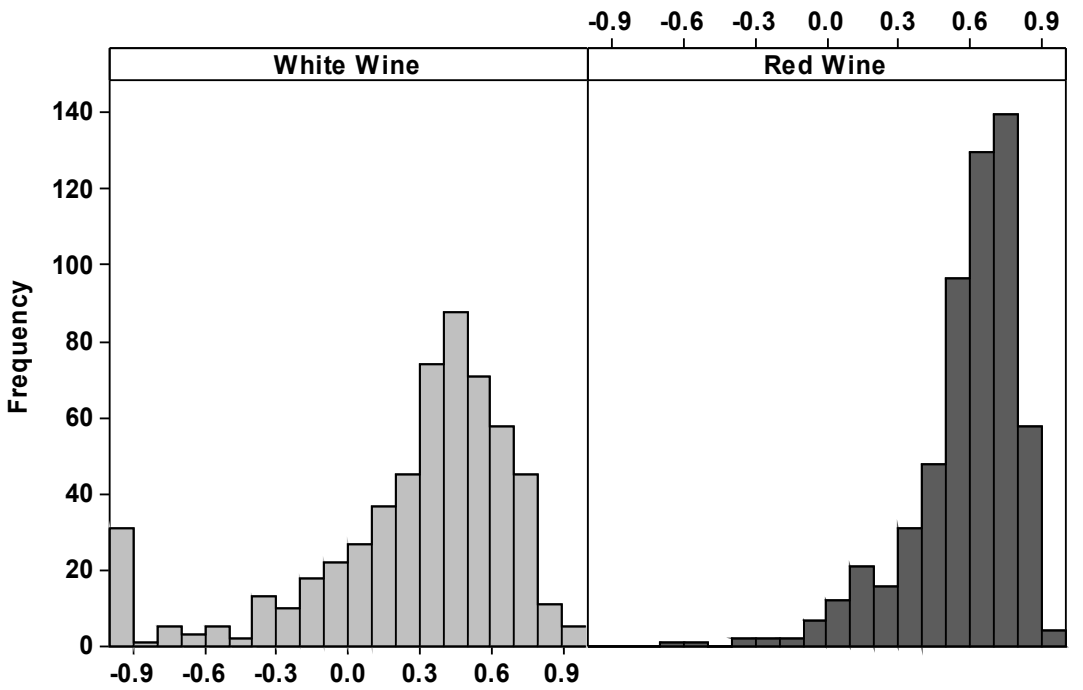

Figure 4a: Distribution of weighted κ statistics for individuals. These represent the assessors ability to consistently classify wines to medal ranges on duplicate presentations (n = 571).The weights are given in Table 2.

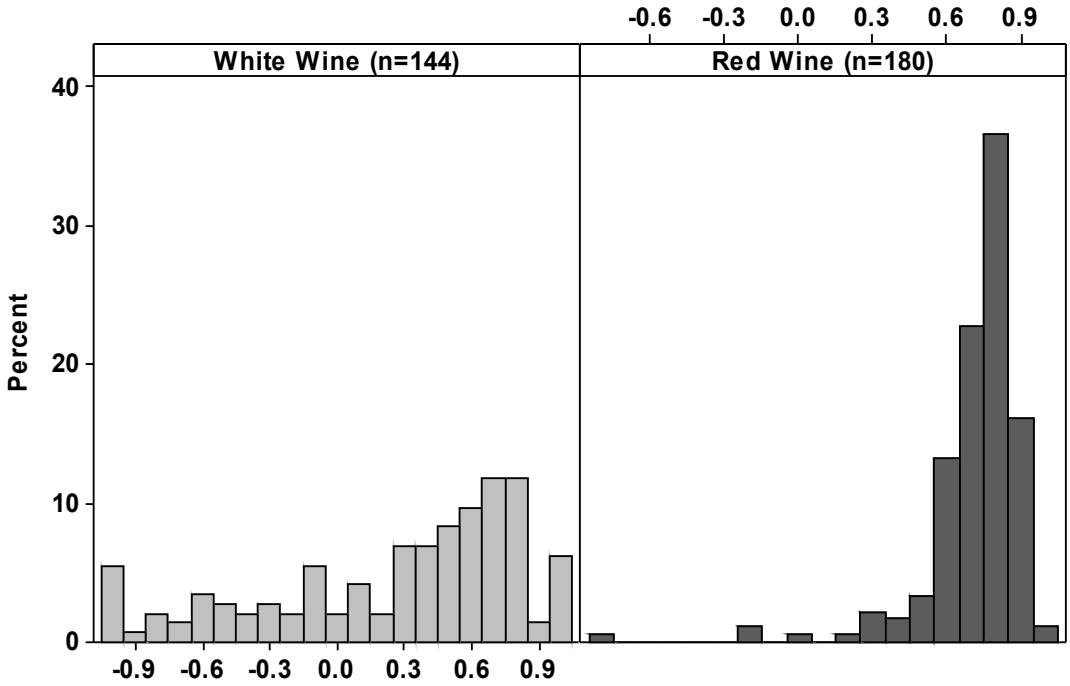

Figure 4b: Distribution of weighted κ statistics for panels of three assessors. These represent the panels ability to consistently classify wines to medal ranges on duplicate presentations. The weights are given in Table 2.