Disponible en ligne sur www.sciencedirect.com
ScienceDirect
Revue européenne de psychologie appliquée 56 (2006) 231–238
Revue européenne de psychologie appliquée
http://france.elsevier.com/direct/ERAP/

Original article

Wine judging, context and New Zealand Sauvignon Blanc

Évaluation de vins, contexte et Sauvignon Blanc Néo-Zélandais

W.V. Parr[a,*], J.A. Green[b], K. Geoffrey White[b]

[a] *Marlborough Wine Research Centre, NZ, Lincoln University, Canterbury, New Zealand*
[b] *University of Otago, Dunedin, New Zealand*

Received 15 March 2005; accepted 5 September 2005

Abstract

Wine show competitions require judges to evaluate a large number of wines, typically within a time constraint. Under such circumstances, some form of quantification of wine quality is essential to achieve the aim of the task, namely allocation of a numerical score, or an award (e.g. a gold medal) that symbolises a numerical range of scores, to each wine. In this paper, we consider the relation between method of scoring, the scores awarded Sauvignon Blanc wines in a simulated wine show competition, and several aspects of wine-judging behaviour. Twenty experienced wine industry professionals judged 15 New Zealand Sauvignon Blanc wines via a 20-point scoring system, the system currently used in New Zealand wine shows, and via a 100-point scoring system in a context that simulated a wine competition. We were interested in two quantitative issues. The first related to the wines, where we investigated whether the 100-point judging system led to greater differentiation of the wines than the 20-point system. The second issue concerned wine-judging behaviour. We were interested in whether scoring method would influence between- and/or within-judge variability, with greater consistency resulting from use of the 20-point system. Results showed that there was no greater differentiation of the Sauvignon Blanc wines when they were judged by the 100-point scale than when judged out of 20 points. Variation in scores given to each wine on each scale was also generally consistent. With respect to whether method of scoring influenced variability of wine judges, we employed a model described by Schlich (1994) to consider measures of wine-evaluation behaviour. The major result was that consistency, both within judges and between judges, was independent of type of scoring method. Implications of the findings are discussed.

Résumé

Lors d'un concours de vins, les juges doivent déguster un grand nombre de vins, généralement dans un temps limité. Dans ces conditions, l'évaluation de la qualité du vin est essentielle pour attribuer à chaque vin, une note ou une distinction (une médaille d'or, par exemple). Dans cet article, nous considérons les relations entre la méthode de notation, les notes obtenues par des vins de cépage Sauvignon Blanc, ainsi que des aspects comportementaux, dans une situation simulée de concours. Vingt experts, professionnels du vin, ont jugé 15 Sauvignons blancs Néo-Zélandais avec un système de notation sur 20 points, actuellement utilisé dans les concours en Nouvelle-Zélande, et un système de notation sur 100 points. Notre questionnement a porté sur deux points. Le premier était relatif aux vins et visait à déterminer si la notation sur 100 était plus discriminante que la notation sur 20. Les résultats ont montré que les vins ne sont pas mieux discriminés avec le système de notation sur 100 points. Le second point était relatif à la variabilité inter- et intra-individuelle. Les résultats ont été analysés selon le model décrit par Schlich (1994). Globalement, la variabilité inter- comme intra-individuelle n'est pas dépendante du système de notation. Les implications de ces résultats sont discutées.

Keywords: Food science; Sauvignon Blanc; Wine judging; Quality; Sensory analysis

Mots clés : Sciences de l'aliment ; Sauvignon Blanc ; Dégustation de vins ; Qualité ; Analyse sensorielle

* Corresponding author. Viticulture and Oenology, Agriculture and Life Sciences, Lincoln University, P.O. Box 84, L.U., Canterbury, New Zealand.
E-mail address: parrw@lincoln.ac.nz (W.V. Parr).

1162-9088/$ - see front matter © 2006 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.erap.2005.09.011

1. Introduction

Assignment of grades, scores, or awards to foods, beverages, and perfumes to reflect 'quality' is a long-established practice, with its roots in industry (trade), rather than in science. Ratings of wine in terms of perceived quality go back at least as far as the mid nineteenth century with events such as the 1855 ranking of the Bordeaux chateaux, and the first Australian Wine Show in 1845 (Walsh, 2002).

Although evaluating wine is a subjective experience that some would prefer not to quantify, standard practice in wine show judging is to allocate a numerical score to each wine. Today, wine competitions are increasingly used by wine producers and wine marketers to take advantage of the opportunities for quality control and advertising that wine shows appear to afford. In Australia and New Zealand, some established wine companies continue to use medals and awards as a dominant factor in their marketing strategies: gold medals do sell wines (Murphy, 2002).

Many wine producers and consumers appear to take a gold medal, or a score of 99 out of 100, at face value. In reality however, there have been few attempts to understand, investigate, and validate the practices involved in formal wine judging events (see Bell, 2003). A few isolated articles have reported data where scientists have considered wine judging behavioural processes such as discrimination, within-judge consistency, and between-judge agreement, using various methods (e.g. Brien et al., 1987; Cliff and King, 1999; Lindley, 2004; Thompson, 2003). However, in the absence of a systematically developed body of knowledge that is easily accessible to wine judges and wine critics, wine professionals tend to rely on anecdotal evidence and their own experience in the wine industry when selecting methods for assessing wine quality.

Wine ratings can take a variety of forms including allocation of stars, a score out of 20 or 100, or allocation of medals. In New Zealand, all formal wine show competitions currently employ a variant of the 20-point scoring system. A 100-point scoring system is often used in several less formal wine tastings, including those where results are published via media such as magazines, internet sites, and newspapers. Some of New Zealand's established wine judges and critics have publicly raised questions concerning the relative effectiveness of different methods for scoring wines. Major points made by wine judges and critics in New Zealand, who are prominent in the wine scoring debate (e.g. Cooper, 2001), centre around whether the method of judging wines influences (i) the scores given to the wines (e.g. does the 100-point scoring system lead to greater differentiation of the wines than the 20-point scale?) and (ii) the consistency of ratings, both between judges and within a judge.

The present study considered the influence of one aspect of the context within which wine judgments were made. Specifically, we investigated wine judging processes in a simulated wine show competition where wine judges evaluated the same Sauvignon Blanc wines out of 20 points (3 = Appearance; 7 = Nose; 10 = Palate) and out of 100 points (15 = Appearance; 35 = Nose; 50 = Palate).

What may at face value appear a relatively trivial issue is sufficiently important to have concerned some New Zealand wine judges, and to have put renowned American and British wine critics on opposite sides of the fence. Robert Parker, whose wine scores exert a powerful economic influence in the United States, and British wine writer Hugh Johnson, have each publicly expressed their support or lack of support, respectively, for the 100- and 20-point scoring systems. Robert Parker defends his use of the 100-point system on his web site where he comments "it is my belief that the various 20-point rating systems do not provide enough flexibility and often result in compressed and inflated wine ratings" (http://www.erobertparker.com/info/legend.asp). Robert Parker further points out that scoring with a 100-point scale in fact begins at 50 points, with each wine gaining a base of 50 points as a result of modern technology. Similarly, the 20-point scale effectively begins at 10, as typically employed in current New Zealand wine competition judging, in that it is unusual for a wine to score below 10.

The 100-point rating scale appears to be most universally accepted in the United States, whereas the 20-point scoring system has a relatively long history in Europe (see Crettenand, 1999), and is currently the dominant method employed in wine shows in New Zealand.

The present paper addressed two questions, one relating to the wines and the other relating to the judges:

- *The wines*: Would the 100-point scoring system lead to greater differentiation of the wines than the 20-point scale?
- *The judges*: Would scoring method influence variability between wine judges and/or variability within a wine judge?

As the research questions were generally based on anecdotal information, we refrained from producing more specific hypotheses. Additionally, we considered performance measures described by Schlich (1994) to assist in interpreting the data, particularly with reference to how individual participants were influenced by scale use.

What the present study did not attempt to do, nor was the design capable of doing, was to investigate the cognitive strategy employed by wine professionals when applying either the 20-point or the 100-point scales. The inherent bias, in terms of greater experience with the 20-point scale than with the 100-point scale, that New Zealand wine judges were likely to exhibit was explicitly acknowledged in the study.

2. Method

2.1. Participants

Twenty wine industry professionals (four-women and 16-men) from Marlborough, New Zealand, experienced at evaluating Sauvignon Blanc wines, participated. Classifying participants on the basis of their experientially gained expertise with the to-be-evaluated product has a precedent in sensory science literature (e.g. Ballester et al., 2005; Bende and Nordin, 1997;

Parr et al., 2002, 2004a). The age range of participants was 26–53 years, with a mean age of 37.2 years. All participants were non-smokers. The mean number of years a participant had spent in the wine industry was 13.7 years (range = 4–26 years), and eight participants had experience in the capacity as a formal wine judge. All 20 participants had previous experience with evaluating wines using a 20-point scale but three people only had previous judging experience with the 100-point scale.

2.2. Materials

The wines were 15 New Zealand Sauvignon Blanc wines that had been previously judged at the 2004 Air New Zealand Wine Awards (ANZWA). In terms of region, 11 wines were from Marlborough, and there was one wine from each of Hawkes Bay, Martinborough, Nelson, and Canterbury. Eleven wines were from the 2004 vintage and four wines were from the 2003 vintage. Of the 15 wines, four had been awarded a silver medal at the 2004 ANZWA competition, seven had been awarded a bronze medal, and four wines had received no award. Two wines were sealed with cork closures whilst the remaining 13 were sealed with screw-cap closures. The wines were coded with numbers, and placed in an order by random selection. Two 750-ml bottles of each wine were used per 8-hour experimental day. Each bottle was temporarily re-sealed between successive pourings of a wine across the 8-hour period.

2.3. Task environment

The study was conducted at the Sensory Facilities of the Marlborough Wine Research Centre. The environment was controlled as advised for sensory laboratories (ASTM, 1986) and International Wine Competitions (O.I.V., 1994). There was a uniform source of lighting, absence of noise and distracting stimuli, and ambient temperature was 18–21.4 °C across the day. Fresh water and small pieces of bread were provided.

2.4. Procedure

The experiment was conducted over two 8-hour days, with 10 judges participating each day. Two judges participated at any one time in five successive sessions, with each session being approximately 11/2 hours in duration. Judges were seated at individual tables covered with white tablecloths on which the wines were served blind in standard, clear glasses. The context simulated a wine show judging. Each judge had been advised at the time that they were invited to participate in the study that they would be taking part in a research study rather than a standard wine tasting or judging. In keeping with ethical agreements, judges were provided with written information about the study and signed a consent form prior to their participation.

Wines were served at ambient temperature. A 50 ml sample of a 2003 Marlborough Sauvignon Blanc that was not employed in the experiment proper was used to condition the palate of each person before they participated in the experiment proper. Each participant rated two identical flights, each comprising 18 Sauvignon Blanc wines. A participant received the 18 wines in the same order across their two flights. One flight of 18 wines was rated using a 20-point scoring system and the other flight was rated using a 100-point scoring system. The 18 wines within a flight comprised 15 unique wines, along with three replicates. Replicated samples were from the wines at positions 1, 8 and 15 of the flight and were presented as wines 16, 17, and 18, respectively.

Participants were instructed to score each wine of a flight individually, in the order set out on the table. Re-evaluation of a wine was permitted once all 18 wines had been scored in their correct order. Odd-numbered participants judged the wines in the randomly selected order, whilst even-numbered participants judged the wines in the reverse of this order. Order of scale use (20-point first or 100-point first) was controlled across judges. Participants recorded their judgments to each wine in writing on data sheets that simulated the score sheets employed in several major wine competitions in New Zealand (e.g. the NZ Wine Society Royal Easter Show). The data sheet provided a column where participants could report salient characters associated with their olfactory and palate (taste and trigeminal nerve stimulation) judgments to each of the 15 New Zealand Sauvignon Blanc wines. These qualitative data, along with the data from the intervening task described below, are reported in detail elsewhere (Parr et al., 2005).

Judging the two flights of wines was separated by a temporal gap of approximately 20 min. During this time a participant engaged in two verbal tasks, a descriptive task and a concept-rating task involving three Sauvignon Blanc wines that were not part of the wine-judging experiment. These intervening tasks served two purposes. First, the verbal tasks were aimed at use of cognitive processes, such as covert language skills, that could substitute for the discussions that typically occur between judging flights of wines in show competitions. Second, the intervening tasks served to minimise memory effects that could be carried over from the first judging to the second flight.

3. Results

All scores out of 20 were converted to scores out of 100 (multiplied by five) to facilitate analysis, but scores out of 20 are presented here to aid interpretation. With respect to the first question, namely that concerning differentiation of the 15 wines, a strong correlation was found between mean scores on the 20- and 100-point scales for each wine, $r\ (15) = 0.87$, $P < 0.001$ (left-hand panel of Fig. 1). The diagonal line represents the equivalent scores on the two scales (i.e. 20-point = 100-point). Overall, wines were awarded similar average scores regardless of scale, and hence, were ranked similarly on average. Greater differentiation on one scale or the other would manifest as the points falling in a line less parallel to the marked diagonal than is observed. Similarly,

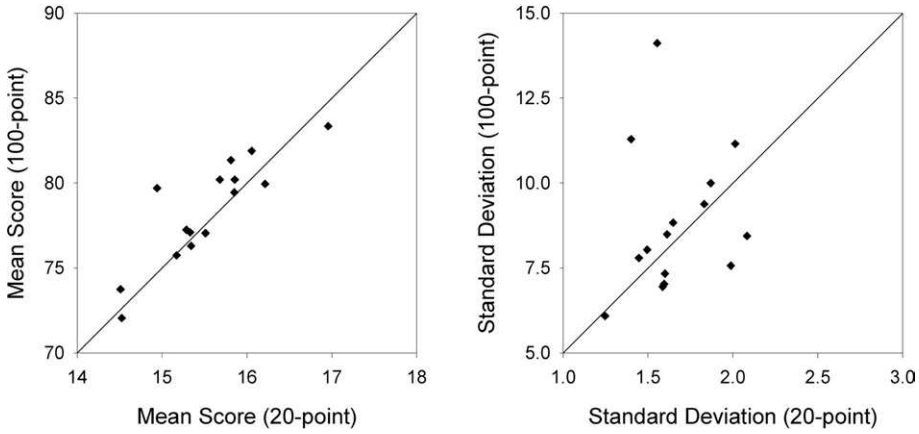

Fig. 1. Mean scores awarded to each wine (averaged across participants) on 20-point versus 100-point scales (left-hand panel), and standard deviation amongst participants' scores for each wine on 20-point versus 100-point scales (right-hand panel). Diagonal lines represent equivalence between scales (i.e. 20-point = 100-point).

inflated scores on one scale or the other would appear as points falling more on one side of the diagonal line than the other.

The right-hand panel of Fig. 1 shows the relationship between the variance amongst participants' scores for each wine on the scales. The relationship was not statistically significant, r (15) = 0.17, ns, but with the exception of two outliers in the top left, most of the wines were scored with similar amounts of variance amongst participants on each scale (as indicated by points falling close to the diagonal). While these outliers might indicate a greater level of variance in participants' scores for the wines on the 100-point scale, the difference in mean standard deviation was not significant, t (14) = 0.19, ns. We will further consider the influence of scales on differentiation of the wines when examining individual differences among participants.

To consider now whether scoring method influenced variability between and within participants, Fig. 2 is identical to Fig. 1, except that data-points now reflect participants rather than wines. The left-hand panel shows that some participants moved their average scores lower or higher on one scale than the other, as signified by points falling further from the diagonal line. Participants' mean scores across wines on the 20-point scale were not significantly related to their mean score across wines on the 100-point scale, r (20) = 0.25, ns. In contrast, the right-hand panel shows that participants who gave a wide spread (large standard deviation) of scores to wines on the 100-point scale, also gave a wide spread on the 20-point scale, r (20) = 0.72, $P < 0.001$.

To consider in more detail how participants were influenced by the scales, we turn to measures outlined in Schlich (1994). Relevant to the present study are his measures of Drift and Discrimination. Conceptually, Drift is the extent to which participants' scores alter in a consistent direction between sessions. As applied to the current work, Drift is the extent to which participants gave higher or lower scores on the 20- versus 100-point scale. However, the measure of Drift is more sophisticated than simply considering the distance of each point from the diagonal line in the left-hand panel of Fig. 2. This is because Drift is a ratio of this distance to the variability of scores awarded by each participant. That is, if a participant awards scores in a very tight range, the same mean movement reflects a far greater change than in a participant who awards wines scores over a much wider range of the scale. In a minor modification to the calculation outlined in Schlich (1994), we

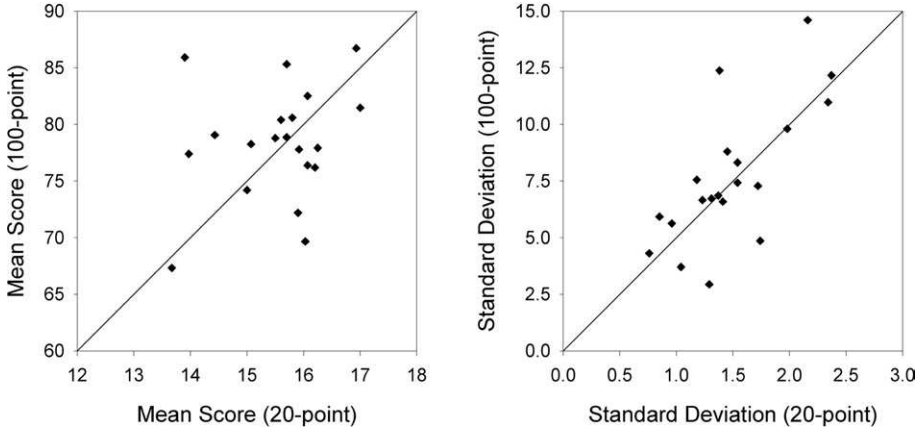

Fig. 2. Mean scores awarded by each participant (averaged across wines) on 20-point versus 100-point scales (left-hand panel), and standard deviation amongst wine scores for each participant on 20-point versus 100-point scales (right-hand panel). Diagonal lines represent equivalence between scales (i.e. 20-point = 100-point).

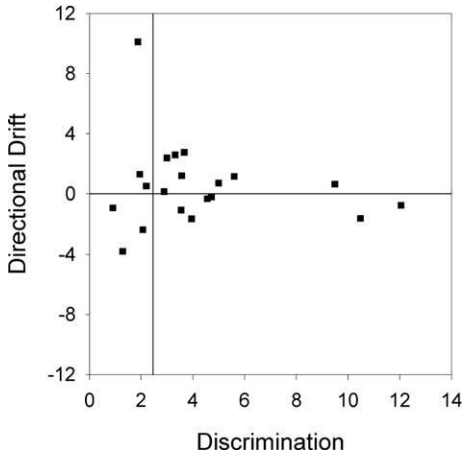

Fig. 3. Graph of Discrimination against Directional Drift for each participant. The vertical reference line represents $F_{critical}$ for Discrimination.

calculated a 'Directional Drift', such that positive numbers indicated higher mean scores on the 100-point scale, and negative numbers indicate lower mean scores on the 20-point scale. Thus, absolute values of those presented here would be identical to those typically calculated.

The measure of Discrimination also taps an important dimension of consistency across scales for participants. In many respects, it is analogous to producing a correlation between each participant's scores on the 20- and 100-point scales. Discrimination is the extent to which participants give each wine a similar score on each scale, controlling for Drift. That is, a participant with a high Discrimination score would give a high score on the 20-point scale to wines they gave a high score on the 100-point scale, and equivalently low scores on both scales to less-preferred wines[1]. Fig. 3 plots each participant's Drift score against their Discrimination score. As Discrimination is an F ratio, it is possible to consider those participants to the right of the vertical reference line ($F_{critical}$) as showing statistically significant discrimination. Fig. 3 shows the number of participants who drifted in either direction was not high, and further, the direction of the drift was not consistent. That is, neither scale lead to any number of participants inflating their scores in a consistent direction. Further, the participants who had the highest levels of drift (in either direction) were also the least likely to show significant discrimination between the different wine samples. While the Discrimination measure is independent of Drift from a computational perspective, it is interesting that those who rated wines similarly across scales (high Discrimination) were also the least likely to shift their anchoring point between scales (Drift).

While the results thus far show no indication of any consistent effect of scale, it is possible that a clearer result has been obscured by variability caused by high levels of Drift, or poor Discrimination by some participants. Thus, the left-hand panel of Fig. 4 depicts the scores for individual wines for the three most discriminating judges, who are clear outliers in terms of high Discrimination in Fig. 3. These participants gave similar scores to each wine regardless of which scale they were presented with, as evidenced by the tight scatter of points around the diagonal equivalence line. In contrast, the right-hand panel of Fig. 4 is a similar graph, but presents the data for the six least discriminating participants. In addition to these participants showing the greatest levels of Drift, it is clear from the roughly random scatter that wines were not receiving equivalent scores on both scales, and that this was not as a result of any systematic change. This shows a great deal of between-subject variability, and with respect to the poorest discriminators, a great deal of within-subject variability.

A possible criticism of our methodology is that participants, being more familiar with the 20-point scale, could simply have multiplied a 20-point score to produce their 100-point score. To consider this, the number of 100-point scores that were multiples-of-five (e.g. 85, 90) was identified. One in five scores would be expected to be such a number, if scores were not biased towards multiples-of-five. Then, using a binomial distribution, it was determined that a participant would have to give more than six scores that were multiples-of-five to be using a significantly different proportion of multiples-of-five scores than expected. When the preceding analyses were re-run including only the 12 participants who did not use a greater than expected number of multiples-of-five (mean number 4.5), the results were unchanged. Further, participants who could be conceived as having converted their 20-point scores to 100 (i.e. used more multiples-of-five) were among the least discriminating, including four of the six worst performing, and not one of the top three. None-the-less, it is still possible that the participants who did not use a large number of multiples-of-five still scaffolded their judgements around these intervals. The present study's design does not permit us to completely rule out this possibility.

What we can conclude, at least with respect to the present judges who were more familiar with the 20-point scale, is that the 100-point scale's greater degree of precision did not lead to greatly different judgements from the 20-point scale.

3.1. Qualitative data

Nineteen of the 20 participants provided descriptors to each of the 15 wines. Participants had been invited to provide 3–4 salient descriptors to each wine (i.e. to report only those characters that appeared dominant during their judging of any wine). Table 1 has been reproduced with kind permission of the *Australian & New Zealand Grapegrower & Winemaker*, in which this information was first published.

Table 1 shows frequencies of descriptor groupings as a function of each of the 15 wines. The wines are ordered from left to right in terms of average score awarded (averaged across scale) in the judging task. The upper quartile (highest 25% by frequency) for each descriptor grouping is bolded, with fre-

[1] For computational details for Drift and Discrimination, see Schlich (1994). To calculate Directional Drift, a negative sign was added to Drift scores where participants mean score was lower on the 20-point scale than the 100-point scale.

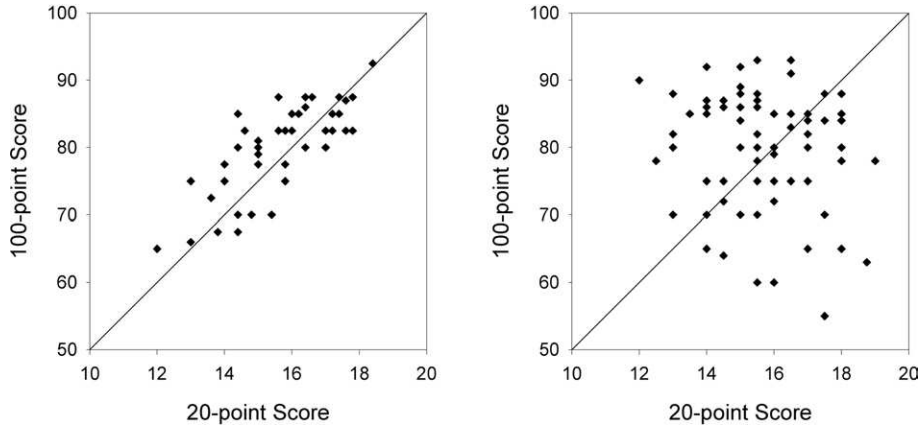

Fig. 4. Scores for each distinct wine for each of the three most discriminating participants as a function of scale (left-hand panel). Scores for each distinct wine for participants with non-significant Discrimination scores (right-hand panel).

Table 1
Frequency of descriptors ($N = 19$ participants) for each wine. Wines are ordered from lowest to highest on the basis of their mean score across scale. Descriptor groupings are ordered by greatest to least-frequent use averaged across wines. Bolded figures represent the upper quartile of frequencies for each descriptor grouping. Zeros are not included, and frequencies less or equal to two are printed in a lighter font

	Mean score for each wine (lowest to highest)														
	72.3	73.2	75.8	76.5	76.8	76.9	77.2	77.3	79.3	79.4	79.7	80.2	80.5	81.1	84.1
Descriptor grouping	Frequency														
Vegetal/herbaceous/grassy/capsicum/green	5	**22**	8	9	**19**	3	9	6	11	**12**	10	11	6	11	11
Tropical/ripe fruits	**10**	3	**10**	7	4	3	**10**	**8**	7	1	8	3	**10**		5
Passionfruit/grapefruit/citrus	**6**	2	4	**6**	**7**	2	**9**	4	1	3	**6**	**6**	**6**	5	**12**
Developed/oxidised/tired/aged[a]	**10**	4	5	6	**7**	3	**9**	**7**	3		**9**		1	2	1
Boxwood/cat's pee/sweaty[a]	4	1	3		4	3	1	3	8		6	4	**9**	**10**	**10**
Good body/palate weight[a]	2		3	**5**		1	**5**	1	3	2	**7**	2	**7**	2	**17**
Thin palate	1	**11**		**7**	1	2	1	1	**3**	1	**3**	1	**3**		
High acid/steely		1		**7**		3	**4**		1	**4**		2	2	1	**5**
Stonefruit	**2**	1		**4**	**4**		1	1	**2**	**2**	**2**	**2**	**2**	1	1
Bitter/phenolic	1	2	1	**3**	2	2	1		2	3	1		**3**	1	**3**
Good balance	1	1		**4**	**3**		2	1	**7**				1	**3**	2
Earthy/dirty[a]	1	**9**	**3**	**3**	**2**		1	1	**2**	1					
Good length	**3**		**3**		**3**	1			2		1	2	**4**	1	2
H_2S/reduced	1	**4**			1	1		**2**	**2**	**3**		1	**5**		1
Good varietal definition[a]			**2**		**2**			1	**3**				**3**		**7**
Gooseberry	1		1	1	**2**	**2**			1		1	**2**		**3**	
Sweet			**1**	**1**		**1**	**1**	**1**		**1**		**1**			**5**
Creamy/butter/oily[a]	**8**		**3**							**1**					
Floral	**2**	**1**	**3**		**1**	**1**					**2**				**1**
Unripe			**1**	**1**		**1**								**2**	
Cork taint							**5**								
Short length									1		1			1	

[a] Significant correlation between frequency of use of descriptor and mean wine score ($P < 0.05$).

quencies of two and less printed in a lighter font. Significant correlations were found between average scores awarded the wines and frequencies of descriptors. These results are generally visible in the Table as patterns created by the bolded numbers clustering to one side or the other. Highly rated Sauvignon Blanc wines were frequently described as exhibiting "boxwood"/"sweaty"/"cat's pee" characters (r (15) = 0.66, $P = 0.007$), "good varietal definition" (r (15) = 0.53, $P = 0.043$), and "good body"/"palate weight" (r (15) = 0.63, $P = 0.01$). Poorly rated wines were most frequently described as "developed"/"oxidised"/"tired" (r (15) = –0.57, $P = 0.03$), "earthy"/"dirty" (r (15) = –0.61, $P = 0.02$), and "creamy"/"buttery"/"oily" (r (15) = –0.55, $P = 0.03$).

4. Discussion

Despite anecdotal evidence to the contrary, the major result of the present study was that the type of scoring system employed for judging Sauvignon Blanc wines had no significant effect on the average scores allocated to the 15 wines. The data demonstrate that wines were generally allocated similar scores, irrespective of scoring method employed. In other words, there is no evidence from the present data that the wine judges utilised to a significant degree the greater opportunity for precision afforded by the 100-point scale.

Several aspects of wine-judging behaviour were assessed, with the major result being the relative consistency demon-

strated by the wine judges across type of scoring system employed. Generally, participants' mean scores were equivalent on the 20- and 100-point scales. That is, a participant who tended to give scores anchored high on one scale gave high scores on the other scale, and similarly with participants who gave lower scores. While participants may vary in terms of where they locate their scores in terms of generally awarding higher or lower scores, and in terms of how much they spread their scores, the majority of the participants were relatively consistent in terms of how they used each of the two scales.

A limitation of the present study is that it was not designed to be able to tell us anything about how judges went about allocating the 20 or 100 points to each wine, despite this being of interest from a psychological perspective. For example, the study cannot tell us whether a judge made a global judgment, allocating a total mark to a wine, or whether the judge employed the three sub-categories on the tasting sheet (Appearance; Nose; Palate) and summed the marks to provide the final score for a wine. This is a topic for future research to address. None-the-less, what is apparent from the data is that, as described by Parker on his web-site with respect to his method of using the 100-point system, the current participants appear to have used the upper half of the scales only. That is, no wine scored below 10 points or below 50 points on the 20- and 100-point scales, respectively[2].

Comment as to whether participants actually judged a wine out of 100 points or used a 20-point system and multiplied by five can be made based on several aspects of the data. First, if participants were merely awarding the wines a score out of 20, and multiplying it by five, we would expect a greater number of scores that were multiples-of-five. A minority of participants did display this tendency. However, excluding the participants who displayed this tendency had no impact on the analyses presented. Second, these latter participants were amongst the least reliable, whereas if a simple multiplication were underlying our findings of consistency regardless of scale-type, we would expect these participants to be the most reliable. It is possible that the remaining participants may have anchored their judgments around a 20-point score, and subsequently refined their judgment to yield a 100-point score. Alternatively, they may have initially conceived their scores out of 100, without recourse to any reference points.

Although the sensory analyses of descriptive data reported in the present paper are limited in terms of sophistication and detail, results across several studies (Parr et al., 2005; Parr et al., 2004b) suggest that the characters in New Zealand Sauvignon Blanc that appear to be associated with "quality" (e.g. higher ratings in the simulated wine-show), and conversely, least associated with wines with lower scores, are: "sweaty/boxwood/cat's pee", "good varietal definition", and "good body/palate weight". The less preferred wines were more often described as "dirty/earthy", "creamy/buttery/oily", and "developed/oxidised/tired". Precisely what "good varietal definition" means is an empirical question currently under investigation.

[2] With the exception of one judge who considered one of the wines to be corked when awarding their 100-point rating. We can but assume that this is an example of within-judge variability, as they did not come to the same decision when they scored the same wine on the 20-point scale.

5. Conclusion

In conclusion, our findings suggest that the current wine-judging system, at least with respect to New Zealand Sauvignon Blanc, is relatively robust in terms of consistency of scores across the 20-point and 100-point methods, despite the relative inexperience of most New Zealand wine judges with the latter method. However, the present data highlight another issue: If within-judge and between-judge inconsistency are independent of choice of scoring system (20-point vs. 100-point), we need to look beyond the "scoring debate" (Cooper, 2001) for ways to improve quantification of wine quality. Having said that, a prominent British statistician Dennis Lindley, who conducted a Bayesian analysis of the data from one of the world's most famous wine tastings (the 1976 tasting by skilled wine connoisseurs of French and American wines in Paris), comments that consistency among judges and between judges may be ideals to work toward rather than realistic concepts.

Acknowledgments

The research was funded by New Zealand Winegrowers, Marlborough Wine Research Centre, and the Foundation for Research, Science and Technology, NZ (grant UOAX0404). We thank Allied Domecq Wines NZ and NZ Winegrowers for supply of wine, Susan Neighbours and Rob Agnew for assistance in carrying out the study, and Mike Trought, Bob Campbell MW, Michael Cooper, and Terry Dunleavy for assistance and valued comment. Finally, we express our sincere gratitude to members of the Marlborough wine industry without whose participation the research would not be possible.

References

ASTM, 1986. Physical Requirement Guidelines for Sensory Evaluation Laboratories, ASTM STP 913. ASTM Publications, Philadelphia.

Ballester, J., Dacremont, C., Le Fur, Y., Etievant, P., 2005. The role of olfaction in the elaboration and use of the Chardonnay wine concept. Food Quality and Preference 16, 351–359.

Bell, G., 2003. WineSense. ChemoSense 5, 8–9.

Bende, M., Nordin, S., 1997. Perceptual learning in olfaction: professional wine tasters versus controls. Physiology and Behavior 62, 1065–1070.

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

Cliff, M., King, M., 1999. Use of principal component analysis for the evaluation of judge performance at wine competitions. Journal of Wine Research 10, 25–32.

Cooper, M., 2001. Critics make points in scoring debate. Sunday Star Times, 14 January.

Crettenand, J., 1999. Tasting cards in international wine competitions. Journal International des Sciences de la Vigne et du Vin, 99–106 (Special Issue Wine Tasting).

Lindley, D.V., 2004. The analysis of a wine tasting. Liquid assets: the international guide to fine wines. http://www.liquidasset.com.

Murphy, P., 2002. Stakeholder presentation—retailer. ASVO Proceedings: who's running this show? Future directions for the Australian wine show system, 31–32. ASVO, Adelaide.

O.I.V., 1994. Standard on International Wine Competitions. Office International de la Vigne et du Vin, France.

Parr, W.V., Heatherbell, D.A., White, K.G., 2002. Demystifying wine expertise: olfactory threshold, perceptual skill, and semantic memory in expert and novice wine judges. Chemical Senses 27, 747–755.

Parr, W.V., White, K.G., Heatherbell, D., 2004a. The nose knows: influence of colour on perception of wine aroma. Journal of Wine Research 14 (2–3), 79–101.

Parr, W.V., Frost, A., White, K.G., Marfell, J., 2004b. Sensory evaluation of wine: deconstructing the concept of 'Marlborough Sauvignon Blanc'. The Australian & New Zealand Grapegrower & Winemaker: 32nd Annual Technical Issue, 63–69.

Parr, W.V., Green, J.A., White, K.G., 2005. Flavour and aroma of New Zealand Sauvignon Blanc. The Australian & New Zealand Grapegrower & Winemaker: 33rd Annual Technical Issue 497a, 100–108.

Schlich, P., 1994. Grapes: a method and SAS program for graphical representations of assessor performances. Journal of Sensory Studies 9, 157–169.

Thompson, M., 2003. The application of Rasch scaling to wine judging. International Education Journal 4 (3), 201–223.

Walsh, B., 2002. Stakeholder presentation—wine committee, Royal Adelaide Wine Show. ASVO Proceedings: Who's running this show? Future directions for the Australian wine show system, 10–12. ASVO, Adelaide.