Journal of Wine Economics, Volume 13, Number 1, 2018, Pages 83–91
doi:10.1017/jwe.2018.1

Consensus between Ratings of Red Bordeaux Wines by Prominent Critics and Correlations with Prices 2004–2010 and 2011–2016: Ashton Revisited and Expanded*

Marc F. Luxen [a]

Abstract

Wine consumers and producers make decisions partly on ratings of wine critics. Research into reliability (correspondence of repeated ratings of the same wines by one critic) and consensus (correspondence of ratings between critics or competitions) have yielded low estimates. However, Ashton (2013), looking at the consensus among only prominent critics of red Bordeaux, vintages 2004–2010, found a correlation of around 0.60. Here, I revisit these data, and extend the analyses to the years 2011–2016 for the same wines, but with additional new critics. Agreement among the critics ($r = 0.57$) of these new years is comparable to those found by Ashton ($r = 0.60$), with a slight upward trend. Overall, critics agree more about what they do not like. Regarding prices and ratings, wines score below-average ratings when they cost less than 35 euro, and higher ratings between 35 and 100 euro. In wines more than 100 euro there is no correlation between ratings and price. (JEL Classification: C99)

Keywords: Bordeaux quality ratings, wine critic agreement, wine critic consensus, wine quality evaluation, hedonic wine evaluation, rating and cost, wine prices.

I. Introduction

Consumers of wine trust the opinion of wine critics (Ashton, 2016; Storchmann, 2012). However, the expertise of wine experts has been criticized by studies that show lack of consistency in their ratings. In one reliability study, Hodgson (2008) collected the scores of around 70 experts who tasted, without knowing this, the same four wine three times during a flight of 30 wines. The experts were perfectly consistent on only about 18% of the wines, and even then only about wines they did not like. Only about 10% of the judges replicated their own scores of the same wine within a single medal group (say Gold-Gold or Bronze-Bronze). Yet another 10%

*I am indebted to an anonymous reviewer for useful comments.
[a] Satori Research, Pluutwerf 43, Gouda, The Netherlands; e-mail: marcluxen@yahoo.com.

scored the same wine Bronze to Gold. This points to low consistency. In another reliability study (Hodgson, 2009a), 122 experts rated the same wines from 1 to 8 in different competitions. About a third of the judges rated one or more of these wines on different occasions in the range of six categories, and another third rated the same wines in the full range of eight categories. These results suggest low reliability of ratings of wine experts.

In 2009, Hodgson (2009b) looked at reliability within whole competitions (i.e., did the same wines receive a similar total score in different competitions). This should yield higher reliability measures, because within a competition the scores of wines are based on scores of multiple judges. He compared the ratings of 375 wines entered into five separate competitions, looking at bronze, silver, and gold medal winners. Of these 375 wines, 106 wines received a gold medal: this is 35% of the wines. Even with so many gold medals, 75% of the wines receiving a gold medal in one competition received no award at all in any other competition.

In a review of 12 studies, Ashton (2012) found a mean consensus (in terms of a correlation coefficient) of 0.34, which is low. Because of the tremendous variation within and between different studies, Ashton concludes that: "*Overall, little support is found for the idea that experienced wine judges should be regarded as experts*" (p. 70).

However, these studies use ratings of judges with different levels of expertise, and maybe prominent critics do better. In 2013, Ashton published a study in this journal (Ashton, 2013) addressing consensus among prominent wine critics of quality ratings of red Bordeaux wines. Using an existing database available at http://bordoverview.com/ (Bolomey and Van der Put, 2017), he addressed the agreement between quality ratings of Robert Parker, Jancis Robinson, Michel Bettane, and Thierry Desseauve from TASTE, James Suckling, Decanter, and La Revue du Vin de France. Ashton reported correlation coefficients for every pair of critics in the years 2004 to 2010. The average pairwise correlation over these years was about 0.60 (this means that the explained variance of the scores of one critic by the score of another critic is 36%).

Because new data have become available regarding vintages 2011–2016, from additional prominent new wine critics, I extended the analysis of the original findings, which allows for a direct comparison of previous critics of red Bordeaux wines with new prominent critics. In addition, I explored the correlation of ratings and prices.

II. Data and Method

I stayed as close as possible to the methodology of Ashton's study. First, I downloaded the entire database on http://bordoverview.com/ (with the kind permission of owners Bolomey and Van der Put). This database consists of quality ratings of 5,188 Bordeaux wines by Robert Parker, Neal Martin, Jancis Robinson, Tim

Atkin, TASTE, James Suckling, Jeff Leve, Decanter, *La Revue du Vin de France*, Jane Anson, Le Point, *Perswijn*, and René Gabriel.

Step 1: Not all critics used the same scales to rate the wines: some used 1 to 5, some use 1 to 20, some use 0–20, or 80–100, and often critics used a range (e.g., 88–90), or added a plus (+) or a minus (–) to their ratings. I simply removed all "+" and "–" and replaced a range (e.g., 88–90 = 89) with the middle point.

Step 2: To make all scales of all critics comparable, I converted them into Z-scores (standard scores), by subtracting the mean of all scores of each individual critic from each individual wine rating, and dividing this by the standard deviation of the scores of that critic. This transformation preserves all original information.

Step 3: If only one critic rated the wine, I removed that wine (430 wines), leaving 4,758 wines. This is a deviation from Ashton, who used only wines that were rated by *all* critics in his study. This is a defendable and clear choice, but I considered throwing a large number of ratings and critics away from the data was statistically speaking not necessary, because pairwise comparisons are possible as long as a wine is rated by at least two critics.

Step 4: I looked at the distribution of the Z-scores for each critic by plotting a histogram. The score distributions of two critics whose were not nearly normally distributed: Decanter (Shapiro–Wilk test: $p < .001$) and *Revue du Vin de France* (Shapiro–Wilk test: $p < .001$), and I removed these from further analyses.

Finally, *LePoint* yielded only one year of data, and because of this I removed *Le Point* from further analysis.

Step 5: I removed all outliers: very high or low Z-scores larger than three or smaller than minus three. The percentage of wines removed was very small, Robert Parker (0.6%), Neal Martin (0.1%), Jancis Robinson (0.1%), Tim Atkin (0.1%), TASTE (0.3%), James Suckling (0%), Jeff Leve(0%), Jane Anson (0.1%), *Perswijn* (0%), and René Gabriel (0.1).

New in the analyses are Neal Martin, Tim Atkin, Jeff Leve, Jane Anson, *Perswijn* (a leading Dutch wine critic magazine), and René Gabriel. Removed from the analyses are *Revue du Vin* and *Decanter*.

III. Results

First, I tried to replicate the findings of Ashton by calculating the correlations between all pairs of critics over the years 2004–2010. I also calculated the mean correlation of each critic with all others over those years (without Fisher-Z transformation) (Table 1).

The correlations are very similar in both studies. Next, I calculated the correlations between the critics for the new years in the database, that is, 2011–2016 (Table 2).

Table 1
Pairwise Correlations among Ratings of Critics

	Ashton Mean	*Luxen Mean*
Robert Parker–Jancis Robinson	0.45	0.45
Robert Parker–TASTE	0.63	0.60
Robert Parker–James Suckling	0.65	0.68
Jancis Robinson–TASTE	0.53	0.53
Jancis Robinson–James Suckling	0.56	0.52
TASTE–James Suckling	0.61	0.62

These results are comparable with the results found by Ashton; ratings of prominent critics still showed a correlation of around 0.60.

Ashton also compared the mean agreement of ratings for classified growths and non-classified growths, and found that these critics agreed more on class growths (0.61) then non-class growth (0.53), a difference of 0.08. I replicated this finding for the years 2011–2016 with a smaller difference of 0.06 (see Table 3). All pairwise correlations were significant for both classified and non-classified growths.

Again, these results are comparable with the results found by Ashton: critics showed about 5 to 10% higher consensus in ratings of classified growths (with the exception of Jane Anson, whose consensus correlation of ratings of non-classified growths dropped about 30% compared with classified growths, and Tim Atkin, whose non-classified growth scores correlated 10% higher than his classified growth scores).

We now turn to prices. I limited the analysis here to a comparison of the two periods 2004–2010 (Ashton study) and 2011–2016 (this study). For a detailed analysis on the relation of ratings of Robert Parker and Jancis Robinson with-price, corrected for inflation, appellation, left-bank/right-bank and classification, see Ashton (2016). The results are in Table 4.

Overall, the correlations between ratings and price have become slightly higher, but only for James Suckling this change was significant at the 0.01 level. The average correlation of all critics' ratings with prices was 0.49, which means that statistically, 24% ($= 0.49^2$) of the variance in prices was explained by critics' ratings.

Figure 1 shows the mean ratings by all critics together of each wine (as Z-scores), plotted against prices: keep in mind that every data-point is a different wine from a different vintage.

Wines start to score average around €35 (Z-scores have an average of 0), below that price, critics give them below average scores.

Figure 1 shows a different trend before and after a price of around 100 euro. Up to around hundred euro, the higher the price, the higher the rating. When prices are

Table 2

Correlations between the Critics for the New Years in the Database: 2011–2016

	Parker	*Martin*	*Robinson*	*Atkin*	*TASTE*	*Suckling*	*Leve*	*Anson*	*Perswijn*	*Gabriel*	*Mean*
Parker	—										0.57
Martin	0.65	—									0.63
Robinson	0.44	0.56	—								0.50
Atkin	0.48	0.56	0.56	—							0.51
TASTE	0.92	0.68	0.40	0.57	—						0.63
Suckling	0.48	0.69	0.53	0.44	NA	—					0.57
Leve	0.75	0.76	0.53	0.56	0.68	0.79	—				0.66
Anson	0.42	0.52	0.42	0.42	0.50	0.48	0.56	—			0.47
Perswijn	0.51	0.60	0.50	0.51	0.68	0.63	0.68	0.46	—		0.57
Gabriel	0.52	0.64	0.54	0.53	NA	0.51	0.63	0.42	0.57	—	0.54
									Grand mean		**0.57**

Note: All correlations are significant at 0.01 level; NA = not available.

Table 3

Pairwise Correlations among Ratings for Classified and Non-classified Growths

	Classified Growth	*Non-classified Growth*	*Difference*
Robert Parker	0.61	0.54	0.07
Neal Martin	0.64	0.59	0.05
Jancis Robinson	0.52	0.48	0.04
Tim Atkin	0.50	0.55	–0.05
TASTE	0.63	0.56	0.07
James Suckling	0.62	0.57	0.05
Jeff Leve	0.67	0.64	0.03
Jane Anson	0.57	0.39	0.18
Perswijn	0.60	0.50	0.10
René Gabriel	0.59	0.51	0.08
Grand mean	0.59	0.53	0.06

Table 4

Correlations of Critic Ratings with Prices 2004–2010 and 2011–2016

	2004–2010	*2011–2016*	*Difference*
Robert Parker	0.44	0.44	0.00
Neal Martin	0.45	0.49	0.04
Jancis Robinson	0.42	0.49	0.07
Tim Atkin		0.45	
TASTE	0.42	0.55	0.13
James Suckling	0.44	0.55	0.11
Jeff Leve		0.56	
Jane Anson	0.56	0.48	–0.08
Perswijn	0.42	0.48	0.07
René Gabriel	0.42	0.44	0.02
Grand mean correlation	0.44	0.49	0.05
Total number of wines	1,951	1,673	

Figure 1

Mean Rating of All Critics over All Years Related to Prices

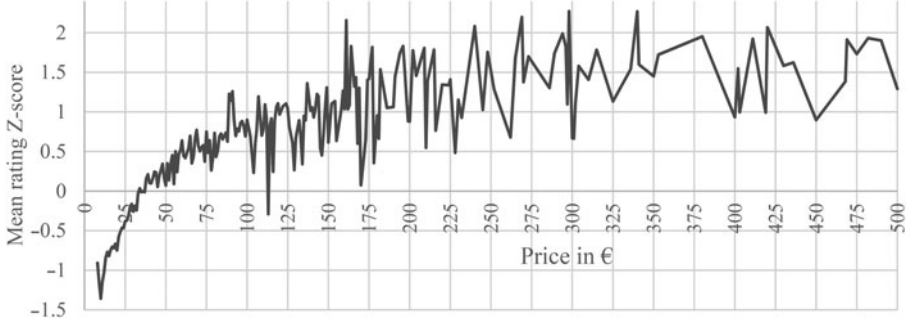

Table 5
Correlations between Ratings and Price Per Price Interval of €50

Price Range	*0–50*	*51–100*	*101–150*	*151–200*	*201–250*	*251–300*
Correlation	0.595	0.303	0.127	0.067	0.111	−0.391
Number of wines	2,628	578	180	69	34	10
Significance	0.00	0.00	0.09	0.59	0.53	0.36

higher than 100 euro, the correlation between rating and price disappears, and the variation becomes large. To formally test this, I calculated correlation coefficients per €50 price interval (Table 5).

As Table 5 shows, ratings and prices were indeed correlated only in wines under around €100. Table 5 also shows that the numbers of wines rapidly diminishes when prices get higher.

It is worth checking the correlation of ratings with price for each critic separately (like in Table 4), but now for wines that cost less than 100 euro. To make comparisons easier, I have also presented the correlations based on the total sample, as in Table 4. The results are reported in Table 6.

The vast majority of wines cost less than 100 euro (1,732 out of 1,951 wines), and this means that sample size was not an issue when selecting those wines. Indeed, when the more expensive wines were removed, the correlations of ratings and price became larger, and sometimes quite substantially so. There were exceptions, however, the rating of Jancis Robinson in 2011–2016 correlate substantially less for wines less than €100 than for all wines together, and the correlations of the ratings of Jane Anson and René Gabriel with price became slightly lower as well.

Table 6
Correlations of Ratings with Price for Each Critic Separately for Wines under €100

	2004–2010			*2011–2016*		
	All Wines	*Wines Less 100*	*Difference*	*All Wines*	*Wines Less 100*	*Difference*
Robert Parker	0.44	0.59	0.15	0.44	0.56	0.12
Neal Martin	0.45	0.60	0.15	0.49	0.51	0.02
Jancis Robin	0.42	0.42	0.00	0.49	0.39	−0.10
Tim Atkin				0.45	0.46	0.01
TASTE	0.42	0.59	0.17	0.55	0.61	0.06
James Suckling	0.44	0.49	0.05	0.55	0.56	0.01
Jeff Leve				0.56	0.63	0.07
Jane Anson	0.56	0.53	−0.03	0.48	0.45	−0.03
Perswijn	0.42	0.54	0.12	0.48	0.58	0.10
René Gabriel	0.42	0.51	0.09	0.44	0.42	−0.02
Grand mean	0.44	0.53	0.09	0.49	0.52	0.03
Total wines	1,951	1,732	−219.00	1,673	1,476	−197.00

IV. Discussion and Conclusion

I examined the level of consensus, or agreement among wine quality ratings of six prominent wine critics for red Bordeaux wines from 2011–2016, and compared them with earlier research by Ashton from the years 2004–2010. The grand mean of consensus across all pairs of critics and all years was 0.57, a figure similar to the one found by Ashton (0.60). Like Ashton, I found that critics agreed more about classified growths (Grand Mean = 0.59; Ashton Grand Mean = 0.63) than non-classified growths (Grand Mean = 0.53; Ashton Grand Mean = 0.51). It seems that our findings are robust. The average explained variance of the rating of a prominent critic by the ratings of the other prominent critics (i.e., the squared correlation) is 35%. This is higher than the explained variance reported in the Ashton overview study (2012) using ratings of wine critics of all levels instead of ratings of prominent critics only like this study of 12% (found by the squaring the correlation of 0.34).

Overall, wines scored below-average ratings when they cost less than 35 euro, and higher ratings when they cost between 35 and 100 euro. There was no correlation between ratings and price in wines that cost more than 100 euro. Most correlations of price with ratings of individual critics get around 0.05 larger when only wines that cost less than 100 euro are considered. This is a common finding: earlier research (e.g., Hodgson, 2009a, 2009b) has shown that agreement between experts about wines they give low scores is higher than for wines they like and give high scores.

There is however a caveat regarding these findings. All these wines were tasted *en primeur* and these tastings are generally not (double) blind. This is a shortcoming in the procedure, because people are sensitive to external cues like price, color, and label when tasting wine. There is no way to know the extent of this effect without additional experiments. On the other hand, end consumers do not buy wines unaware of price either, and this means blind studies are not, and maybe should not be, the golden standard (see Cohen, 2016).

This study shows that consensus among prominent critics, in different constellations over two periods of time, is substantial and stable, which is an important and encouraging finding.

References

Ashton, R. H. (2012). Reliability and consensus of experienced wine judges: Expertise within and between? *Journal of Wine Economics*, 7(1), 70–87.

Ashton, R. H. (2013). Is there consensus among wine quality ratings of prominent critics? An empirical analysis of red Bordeaux, 2004–2010. *Journal of Wine Economics*, 8(2), 225–234.

Ashton, R. H. (2016). The value of expert opinion in the pricing of Bordeaux wine futures. *Journal of Wine Economics*, 11(2), 261–288.

Bolomey, D., and Van der Put, W. (2017). Bordoverview. http://bordoverview.com/ (accessed 10 August 2017).

Cohen, J. (2016). Wine tasting, blind and otherwise: Blindness as a perceptual limitation? https://pdfs.semanticscholar.org/b678/d8b5316a2d43204eb19d7d95cce061700640.pdf (accessed 20 August 2017).
Hodgson, R. T. (2008). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.
Hodgson, R. T. (2009a). An analysis of the concordance among 13 U.S. wine competitions. *Journal of Wine Economics*, 4(1), 1–9.
Hodgson, R. T. (2009b). How expert are "expert" wine judges? *Journal of Wine Economics*, 4(2), 233–241.
Storchmann, K. (2012). Wine economics. *Journal of Wine Economics*, 7(1), 1–33.