Is There Consensus Among Wine Quality Ratings of Prominent Critics? An Empirical Analysis of Red Bordeaux, 2004–2010*

Robert H. Ashton [a]

Abstract

This paper examines the level of consensus, or agreement, among the wine quality ratings of six prominent wine critics for seven consecutive vintages of red Bordeaux. Consensus, a critical component of expertise in wine evaluation, has important implications for consumers' reliance on critics' ratings in deciding which wines to purchase or consume. The principal analyses focus on a core set of wines in each year that were rated by all six critics. Additional analyses concern differences in agreement for classified growths vs. nonclassified growths and for critics of different nationalities (American, British, and French). The level of consensus among these prominent critics is contrasted with that among both wine professionals who are *not* prominent critics and professionals from several other fields. (JEL Classification: C93)

Keywords: Bordeaux quality ratings, wine critic agreement, wine critic consensus, wine quality evaluation.

I. Introduction

Wine consumers often rely on renowned wine critics such as Robert Parker or Jancis Robinson and prominent magazines such as *Wine Spectator* or *Decanter* in deciding which wines to purchase or consume. These and other critics and publications provide assessments of thousands of wines every year in both numerical (rating scales) and verbal (tasting notes) forms. An implicit assumption underlying consumers' reliance on such assessments is that wine critics possess expertise

---

in both the evaluation of wines and the communication of that evaluation in the form of numerical ratings and tasting notes.

Expertise is a difficult concept to define in the domain of wine evaluation. However, as in many areas in which people rely on the assessments of others, consensus—or agreement—has been identified as a critical component of expertise in wine evaluation (e.g., Ashenfelter and Quandt, 1999; Ashton, 2012; Cicchetti, 2004a, 2004b; Hodgson, 2008, 2009a, 2009b). Quandt (2007, 130) makes this point, but more colorfully: "Two things have to be true before wine ratings can become useful for the average wine drinker. Since there are many wine writers, and there is a substantial overlap in the wines they write about (particularly Bordeaux wines), it is important that there be substantial agreement among them. And secondly, what they write must actually convey information; that is to say, it must be free of bullshit. Regrettably, wine evaluations fail on both counts." Quandt goes on to discuss the second requirement while I address the first. Specifically, I provide an empirical analysis of the level of agreement among the ratings of six prominent wine critics in the evaluation of seven consecutive vintages of red Bordeaux.

As Quandt implies, consumers' reliance on critics' ratings is complicated by the existence of multiple critics who rate many of the same wines. To the extent that critics *agree,* it matters little which specific critic a consumer chooses to follow; in fact, agreement among critics will likely serve to increase consumers' confidence in the entire enterprise of wine evaluation. But what should consumers do when faced with *disagreement* among critics, especially critics who have achieved national or international renown? Short of ignoring critics' recommendations altogether, consumers might identify a critic whose "taste profile" is similar to their own and simply follow that critic's advice. Indeed, this strategy has been embraced in a recent consumer-oriented book by Taber (2011) who makes one set of wine recommendations for consumers whose taste profiles are similar to Jancis Robinson's and a very different set of recommendations for consumers with more Robert Parker–like profiles.

Taber's choice of Robinson and Parker as contrasting examples of critics' taste profiles reflects what is widely seen as their divergence in wine appreciation. Never has that divergence been clearer than in their spectacular disagreement concerning the 2003 Château Pavie (e.g., McCoy, 2005; Robinson, 2004; Taber, 2011; Voss, 2004). Parker initially rated the 2003 Pavie 96-100 (later settling on 98 +), describing it as a "stunningly complete wine of irrefutable nobility" [and] "a wine of sublime richness, minerality, delineation . . . ," [with] "provocative aromas" [and] "extraordinary richness as well as remarkable freshness and definition" (eRobertParker.com). In contrast, Robinson rated the 2003 Pavie 12 (on a 10- to 20-point scale), describing it as a "ridiculous wine more reminiscent of a late harvest Zinfandel than a red Bordeaux with its unappetizing green notes" [and] "completely unappetizing overripe aromas" (JancisRobinson.com). Isolated examples like this are instructive, but they do not speak to the more basic issue of how well several prominent critics agree on a large sample of wines across many vintages. It is this more basic issue that I address here.

To my knowledge, this is the first study of agreement among the ratings of several prominent wine critics. Some earlier studies, reviewed by Ashton (2012), have examined consensus among the ratings of experienced wine professionals such as winemakers, wine merchants, wine magazine writers, restaurateurs, sommeliers, and tasters in wine competitions. Each study involved several professionals who, in blind tastings, independently rated a number of wines. Consensus was measured as the correlation across the wines between the ratings of each pair of tasters. Pairwise correlations are often used to measure consensus, in wine studies and many other areas, as correlations are robust with respect to the different rating scales used, and they enhance the comparability of measured agreement levels across individuals and tasks. The average level of consensus across all experienced wine professionals in all studies is .34, indicating that only 12 percent ($.34^2$) of the variability in ratings is common across the wine professionals studied. It may be reasonable to expect prominent wine critics to agree to a greater extent than the broad assortment of wine professionals studied previously. But *how much* closer should we expect their level of consensus to be?

II. Data

The data for this study come from the website bordoverview.com, created and maintained by David Bolomey, a wine merchant/consultant in Amsterdam. This source contains numerical ratings assigned by prominent critics from the United States and Europe for hundreds of red Bordeaux wines from the 2004 to 2010 vintages. The wines were tasted *en primeur* in the spring following the fall harvest and generally about two years before release of the wines to the public. The average number of red Bordeaux in the database is 362 per year, with a range of 245 to 412.

The ratings of six critics who evaluated a large number of wines in each of the seven years are the focus of the present analysis. Two critics are American, two are British, and two are French. These critics are listed in Table 1 along with information about their affiliations and the rating scales that they employ. The average number of wines rated by at least one of these six critics is 355 per year, and the average number rated by any pair of critics is 189 per year. More valuable for current purposes, however, is that an average of 98 wines per year were rated by *all six* critics. The principal analyses in the next section focus on the "core set" of red Bordeaux that were rated by all critics.

III. Results

Pairwise correlations for the core set of wines are shown in Table 2. With six critics, there are 15 possible pairwise correlations in each of the seven years. Five values are missing in 2010 as the BD ratings stop in 2009. All 100 pairwise correlations in Table 2 are positive. In fact, all 100 are significantly different from zero at the .01 level. The grand mean of these pairwise correlations is .60, which is substantially

Table 1
Wine Critics Included in the Study

Robert Parker (RP)
American wine critic who publishes in *The Wine Advocate.* Wines are rated on a scale from 50 to 100.

Jancis Robinson (JR)
English wine critic who publishes on her website JancisRobinson.com. Wines are rated on a scale from 10 to 20.

Michel Bettane & Thierry Desseauve (BD)
French wine critics who publish in *TAST*. Wines are rated on a scale from 10 to 20.

James Suckling (JS)
American wine critic who publishes in *Wine Spectator* up to the 2009 vintage and beginning in 2010 on his website JamesSuckling.com. Wines are rated on a scale from 50 to 100.

***Decanter* (DE)**
English wine magazine. Before 2007, a five-point scale is used. Beginning in 2007, a 10- to 20-point scale is used. The wines are tasted by Steven Spurrier, James Lawther, and Beverley Blanning.

***La Revue du Vin de France* (RVF)**
French wine magazine. Wines are rated on a scale from 10 to 20. The wines are tasted by Olivier Poels, Hélène Durand, and Philippe Maurange.

The above information is an edited version of that provided on bordoverview.com.

Some of the ratings on bordoverview.com are presented as ranges instead of point values: All the JS and RVF ratings are ranges. All the RP 2009 and 2010 ratings are ranges, as are a few RP ratings from earlier vintages. I have simply taken the midpoint of these ranges.

Some critics occasionally append a + or − to their ratings (RP, JR, and DE). I have simply dropped any + or – that appears in the database.

While bordoverview.com states that the RVF ratings are on a scale from 10 to 20, their 2004 and 2005 ratings actually use a five-point scale.

greater than the mean consensus level of .34 for the experienced wine professionals in the studies reviewed by Ashton (2012).

Much variability exists in the mean agreement level of *pairs* of critics, ranging from .45 for Robert Parker (RP) and Jancis Robinson (JR) to .69 for BD and *Decanter* (DE), while there is somewhat less variability in mean agreement across *years*, ranging from .53 to .67. Specific critic pairs agree well in some years and poorly in others, but with no clear pattern over time. The mean correlation of .45 between Robert Parker and Jancis Robinson, which reflects the lowest level of agreement among all pairs of critics, is of particular interest. Further analysis reveals that Robinson agrees relatively little with *all* the other critics, not just Parker. This can be seen by calculating, for each critic, the average of their correlations with the other five critics. The average correlation of Robinson with the other five critics is .52; the average correlations for the other critics range from .60 to .64. I return to this issue in Section IV.

The above analyses were repeated with wines that were rated by *any* pair of critics (mean = 189), as opposed to *all* pairs of critics (mean = 98). Even though the set of wines included in this analysis varies widely across both critic pairs and years, the results are consistent with those in Table 2: The grand mean is .58 (compared to .60 in Table 2) and the other results reported above hold as well.

Table 2
Pairwise Correlations Among Ratings of Six Wine Critics: Core Set of Wines

Critic Pair	2004	2005	2006	2007	2008	2009	2010	Pair Means
RP – JR	.54	.34	.45	.31	.58	.45	.50	.45
RP–BD	.61	.65	.67	.59	.65	.59		.63
RP – JS	.59	.67	.67	.51	.72	.72	.69	.65
RP–DE	.52	.63	.62	.59	.77	.74	.70	.65
RP–RVF	.65	.55	.67	.57	.72	.51	.64	.62
JR–BD	.56	.47	.59	.54	.51	.50		.53
JR – JS	.65	.51	.52	.46	.75	.46	.55	.56
JR–DE	.54	.55	.57	.52	.60	.57	.70	.58
JR–RVF	.55	.29	.62	.37	.53	.49	.57	.49
BD–JS	.55	.71	.67	.63	.55	.57		.61
BD–DE	.60	.77	.66	.63	.78	.69		.69
BD–RVF	.58	.63	.70	.63	.73	.60		.65
JS–DE	.56	.65	.61	.46	.67	.64	.72	.62
JS–RVF	.63	.51	.66	.48	.76	.55	.65	.61
DE–RVF	.53	.56	.67	.66	.81	.68	.79	.67
Year Means	.58	.57	.62	.53	.67	.58	.65	.60

RP = Robert Parker
JR = Jancis Robinson
BD = Michel Bettane & Thierry Desseauve
JS = James Suckling
DE = *Decanter*
RVF = *La Revue du Vin de France*

While Table 2 presents pairwise consensus measures for all wines in the core set (mean = 98), Tables 3 and 4 disaggregate the results for classified growths (mean = 59) and nonclassified growths (mean = 39), respectively. Classified growths include wines in the five levels of the Medoc *Grand Cru* classification, as well as the Saint-Emilion *Premier Grand Cru Classe A* and *B* and the classified Graves from Pessac-Leognan, consistent with the definition of "classified growths" employed by Hadj Ali and Nauges (2007). Prior research showing that experienced wine professionals agree better for wines of higher quality (Cliff and King, 1997; Hodgson, 2008, 2009a) suggests that the critics' ratings for the classified growths will agree better than their ratings for the nonclassified growths.

This is indeed the case. The grand mean of the 100 pairwise correlations for the classified (nonclassified) growths is .63 (.51). Moreover, the mean correlations across years are greater for classified growths than for nonclassified growths for all 15 pairs of critics, and the mean correlations across critic pairs are greater for classified growths than for nonclassified growths for all seven years. A further indication of the greater consensus for classified growths is that 99 of the 100 correlations in Table 3 are significant at the .01 level, while only 62 of the 100 correlations in Table 4 are significant at .01.

Thus, it is clear that, on average, consensus among these critics is greater for the classified growths than for the nonclassified growths. But are there pairs of critics

Table 3
Pairwise Correlations Among Ratings of Six Wine Critics: Classified Growths

Critic Pair	2004	2005	2006	2007	2008	2009	2010	Pair Means
RP–JR	.58	.34	.41	.35	.58	.46	.51	.46
RP–BD	.61	.68	.69	.71	.61	.59		.65
RP–JS	.56	.67	.72	.59	.74	.78	.67	.68
RP–DE	.49	.67	.69	.72	.74	.76	.69	.68
RP–RVF	.64	.59	.64	.79	.70	.58	.72	.67
JR–BD	.62	.44	.57	.57	.60	.54		.56
JR–JS	.67	.42	.45	.55	.76	.44	.54	.55
JR–DE	.57	.50	.63	.57	.73	.61	.70	.62
JR–RVF	.62	.22	.63	.42	.56	.56	.56	.51
BD–JS	.68	.74	.74	.66	.69	.64		.69
BD–DE	.59	.76	.72	.69	.75	.65		.69
BD–RVF	.66	.71	.70	.73	.70	.59		.68
JS–DE	.61	.70	.75	.58	.77	.69	.74	.69
JS–RVF	.67	.54	.67	.53	.82	.66	.75	.66
DE–RVF	.52	.53	.74	.78	.82	.65	.80	.69
Year Means	.61	.57	.65	.62	.70	.61	.67	.63

RP = Robert Parker
JR = Jancis Robinson
BD = Michel Bettane & Thierry Desseauve
JS = James Suckling
DE = *Decanter*
RVF = *La Revue du Vin de France*

who agree to a greater extent on the nonclassified growths? This can be determined by comparing each of the 100 pairwise measures in Table 3 with its counterpart in Table 4. Twenty-five such values are greater in Table 4, implying greater consensus for the nonclassified growths, but none of the differences is significant even at .05. In contrast, 75 values are greater in Table 3, and 21 of these differences are significant (17 at .05; four at .01). As with the overall analysis, however, there is no clear pattern across critic pairs.

Substantial pairwise variability in consensus is evident for both the classified growths and the nonclassified growths. Another result that is evident for both subsets of wines is the relatively low agreement between Robert Parker and Jancis Robinson. These two critics produce the lowest mean correlation across years for both classified growths (.46) and nonclassified growths (.41). Moreover, Robinson is again found to agree relatively poorly with *all five* of the other critics. For the classified growths, her average correlation with the other five critics is .54 compared to a range of .63 to .67 for the others. For the nonclassified growths, her average is .45 compared to a range of .51 to .54. I return to this issue in Section IV.

The above analyses were repeated with Château Petrus included with the classified growths. Petrus, a Pomerol, is widely considered on a par with first-tier Medocs even though there has never been an official classification of Pomerols. The analyses were repeated again with Petrus and nine *additional* Pomerols included

Table 4
Pairwise Correlations Among Ratings of Six Wine Critics: Nonclassified Growths

Critic Pair	2004	2005	2006	2007	2008	2009	2010	Pair Means
RP–JR	.41	.33	.45	.20	.64	.39	.46	.41
RP–BD	.61	.62	.59	.40	.74	.50		.58
RP–JS	.68	.66	.53	.35	.66	.63	.70	.60
RP–DE	.56	.59	.41	.35	.88	.69	.71	.60
RP–RVF	.66	.52	.72	.17	.79	.28	.52	.52
JR–BD	.41	.46	.57	.47	.19	.40		.42
JR–JS	.62	.58	.57	.22	.74	.48	.55	.54
JR–DE	.45	.57	.37	.38	.22	.48	.67	.45
JR–RVF	.43	.38	.60	.23	.41	.30	.52	.41
BD–JS	.37	.67	.53	.59	.12	.46		.46
BD–DE	.61	.75	.51	.34	.85	.56		.60
BD–RVF	.45	.50	.71	.27	.82	.39		.52
JS–DE	.42	.57	.33	.13	.43	.62	.73	.46
JS–RVF	.61	.48	.66	.30	.55	.39	.57	.51
DE–RVF	.54	.65	.49	.31	.84	.47	.70	.57
Year Means	.52	.56	.53	.31	.59	.47	.61	.51

RP = Robert Parker
JR = Jancis Robinson
BD = Michel Bettane & Thierry Desseauve
JS = James Suckling
DE = *Decanter*
RVF = *La Revue du Vin de France*

with the classified growths—Clinet, Clos l'Eglise, Hosanna, la Fleur-Petrus, la Conseillante, la Violette, l'Eglise-Clinet, l'Evangile, and le Pin. Inclusion of these additional Pomerols (which is admittedly somewhat ad hoc) is based on the favorable press and ratings that they have garnered in recent years. Neither of these alternative definitions of "classified growths" produces results that differ from those reported in Tables 3 and 4. For example, the grand mean of consensus (.63 per Table 3) is .64 if only Petrus is included with the classified growths and .61 if Petrus plus the other nine are included.

Finally, it may be of interest to examine the level of consensus between the two American critics (RP and JS), the two British critics (JR and DE), and the two French critics (BD and RVF) vis-à-vis critic pairs of different nationalities. It is sometimes remarked that when controversies such as that involving the 2003 Pavie arise, British critics tend to agree with Robinson while American critics tend to agree with Parker, an alignment often ascribed to critics' preferences for "elegance" vs. "power." One might also wonder whether French critics will agree to a greater extent on French wines than will American or British critics.

Tables 2–4 reveal above-average agreement for the two American and the two French critics and below-average agreement for the two British critics. In Table 2, mean agreement for both the American and French critic pairs is .65 but is .58 for the British pair. An almost identical pattern is apparent in Table 3

(classified growths), while a somewhat greater discrepancy in favor of the Americans and the French is apparent in Table 4 (nonclassified growths). The differences are not particularly large, but they do indicate greater consensus for the Americans and the French than for the British. However, since one of the British critics (Robinson) agrees somewhat less with *all* the other critics, not just *Decanter*, any interpretation of these "national differences" must keep that in mind.

IV. Discussion

In this paper I examine the level of consensus, or agreement, among wine quality ratings of six prominent wine critics for red Bordeaux wines from 2004 to 2010. Quantifying consensus as the correlation across wines between the ratings of each pair of critics, it is found that, for the overall set of wines, all pairwise correlations are significantly positive for all years. The grand mean of consensus across all pairs of critics and all years is .60. Disaggregating the overall set of wines into classified growths and nonclassified growths reveals greater consensus for the former (grand mean = .63) than for the latter (grand mean = .51), as well as a decrease in the number of pairwise correlations that reach significance for the nonclassified growths.

Given these results, one might ask whether the glass is half-full or half-empty. How does one decide whether mean consensus of .60 among such renowned critics is good or bad? A glass-half-*full* interpretation would observe that mean consensus in earlier studies of experienced wine professionals who are *not* prominent critics has been documented as only .34 (Ashton, 2012). Therefore, the average amount of pairwise common variance in the ratings of these wine professionals is 12 percent ($.34^2$) while that of the six prominent critics is 36 percent ($.60^2$), a threefold increase. In addition, there are numerous instances of *negative* correlations among pairs of tasters in the earlier studies, but there are no such instances among these prominent critics, for either the classified growths or the nonclassified growths. Thus, these six critics clearly agree more than the wine professionals studied earlier and, to the extent that agreement signals expertise, they can be considered "more expert."

Another relevant point of comparison is the mean level of consensus found in fields other than wine evaluation. Ashton (2012) summarizes the results of 46 consensus studies conducted across six professional fields. The fields and their mean consensus levels are: meteorology (.75), personnel management (.65), auditing (.61), medicine (.56), business (.49), and clinical psychology (.37). Considering the subjective nature of wine evaluation, a mean consensus level of .60 among prominent wine critics may be viewed quite favorably.

However, a glass-half-*empty* interpretation would observe that a mean pairwise correlation of .60 leaves almost two-thirds of the variability in the critics' ratings unexplained. This level of agreement might be considered low in a setting where

several factors that could be expected to dampen agreement levels are controlled; in particular, all the wines in the present analysis are red Bordeaux, in contrast to the studies reviewed by Ashton (2012), in which the wines rated were often from different countries and different grape varieties, even of different colors. It is also important to note that Bordeaux *en primeur* tastings are generally not blind, in contrast to the tastings done by the experienced professionals in the earlier studies. Clearly, this fundamental difference in the conduct of the tastings favors the wine critics (as does the fact that they tasted only red Bordeaux) but the extent of this "agreement advantage" cannot be estimated given the results that are available.

Not surprisingly, some pairs of critics tend to agree better than others. However, there is no clear pattern in relative agreement across either critic pairs or years. The only exception concerns Jancis Robinson, whose ratings are, on average, somewhat "out of line" with those of the other five critics. Her ratings are *most* out of line with those of Robert Parker. These two critics produce the lowest mean agreement level of all 15 critic pairs—overall and for both the classified growths and nonclassified growths. This finding lends credibility to statements that wine writers often make about the different "tastes" of Robinson and Parker (e.g., Burnham and Skilleas, 2008; Feiring, 2008; Taber, 2011; Voss, 2004). Taber (2011, 38), commenting on this issue, says that "One is not wrong, and the other is not right. They're simply different, in exactly the same way that some people like the music of Brahms and others prefer Copland."

It is, of course, the *dis*agreement—not the agreement—between the tastes (and ratings) of prominent critics that attracts the greatest attention from both consumers and commentators. As mentioned earlier, the disagreement between Parker and Robinson regarding the 2003 Château Pavie attracted substantial attention. In contrast, neither the agreement between Parker and *other* critics on the 2003 Pavie nor the agreement between Parker and Robinson herself on subsequent Pavie vintages seems to have attracted any attention whatsoever. And 30 years ago it was the disagreement between the newcomer Robert Parker and more established critics on the 1982 Bordeaux vintage that launched Parker's career as the world's most influential wine critic. Whatever the field, one is unlikely to become known as an expert among experts by agreeing with everyone else.

In an odd twist to the Pavie saga, the critic John Gilman, himself a newcomer who publishes *View from the Cellar*, a web-only publication, has recently offered an extremely negative evaluation of the 2010 Pavie, in stark contrast to the glowingly positive evaluations by Parker and other established critics, including Robinson (Asimov, 2012). Parker, for example, rated the 2010 Pavie 95–98 +, describing it as possessing "full-bodied power and sensational density, texture and length." Gilman rated it 47–52 +, describing it as "absurdly overripe, unpleasant to taste and patently out of balance" [and] "the biggest train wreck of the vintage" (Asimov, 2012). Perhaps Gilman's view of the 2010 Pavie will eventually prevail, but in the meantime Château Pavie, along with Château Angelus, was recently promoted from *Premier Grand Cru Classe B* to *Premier Grand Cru Classe A*