AN ANALYSIS OF WINE CRITIC
CONSENSUS: A STUDY OF WASHINGTON
AND CALIFORNIA WINES

Eric T. Stuen, Jon R. Miller and
Robert W. Stone
An Analysis of Wine Critic Consensus:
A Study of Washington and California Wines

Eric T. Stuen*
Jon R. Miller*
Robert W. Stone*

June 2014
Preliminary and Incomplete

Abstract

We examine the degree of consensus in quality ratings of US among prominent wine publications. Ratings are an important source of information for both wine consumers and wine researchers. For the purpose of wine research, are ratings on the ubiquitous 100 point scale reliable, objective measures of quality? The value of expert judgment has been called into question by a number of studies, especially in the context of wine competitions and tasting events. Our study is part of a much smaller literature focusing on ratings by expert critics. We look at four publications: Wine Spectator (WS) and Wine Enthusiast (WE), which review a broad selection of the wine market, and Wine Advocate (WA) and International Wine Cellar (IWC), which are more selective and focus more on the high-end of the market. We find a similar level of consensus, measured by the correlation coefficient, between some pairs of critics regarding wines from California and Washington as Ashton (2013) does for critics of Bordeaux wine. However, among other pairs the correlation is much lower, suggesting almost no consensus. Consensus is not found to be related to the blinding policies (or lack thereof) of the critical publications. Our findings show that quality ratings have a substantial degree of objectivity to them.

*Stuen, Miller and Stone are, respectively, Assistant Professor of Economics, Professor of Economics, and Professor of Accounting in the College of Business and Economics, University of Idaho.

We thank Dominic Cicchetti, Omer Gokcekus, Robin Goldstein, Lynne Stokes and other participants at the 2014 AAWE Annual Conference for helpful comments, and thank Stefan Richardson for excellent research assistance.
An Analysis of Wine Critic Consensus: A Study of Washington and California Wines

INTRODUCTION

Due to strong information asymmetries in wine markets, wineries and wine retailers often use expert ratings or the results of competitive awards to signal wine quality. If the ratings of wine critics truly measure wine quality, consumers benefit by being assured of getting what they pay for. There is consistent evidence in the literature that expert ratings are positively related to wine prices, which should imply that consumers who pay high prices for wines receive commensurate benefit. However, a large and growing literature has shown that consumers do not necessarily prefer more expensive and “reserve” wines in blind tastings (Weil, 2005; Goldstein et al., 2008) and that there is a high degree of inconsistency in wine competitions, both between judges in a particular competition, but especially between competitions (e.g., Hodgson, 2009a). Taken together, these studies suggest that consumers may be misled by competition awards and critical ratings.

A second reason the consistency of wine ratings is important pertains to wine economics. Hedonic wine pricing studies, as well as studies of signaling behavior, depend on accurate measures of wine quality. Without such knowledge, the effects of quality and other factors on market behavior may be miss-estimated. Therefore, it is necessary to evaluate the extent to which wine ratings are objective measures of quality.

In this paper, we extend the limited literature on wine critic consensus to include analysis of U.S. wines from the largest wine producing states, California and Washington. Previous studies in this area focused on Bordeaux wines, showing that professional wine critic ratings have strong concordance with both online community ratings (Gokcekus and Nottebaum, 2011) and with each other (Ashton, 2013). We assemble and compare wine critic ratings from four wine publications, *Wine Spectator* (WS), *Wine Enthusiast* (WE), *The
Wine Advocate (WA), and International Wine Cellar (IWC), to undertake a similar analysis for California and Washington wine.

We focus on consensus between wine critics, measured primarily by inter-critic correlation, because “true” quality values are unknown and because repeated tastings of the same wine by a critic (which would establish intra-critic correlation, or reliability) are seldom done. Ashton (2012) reports on the mean correlation coefficients across studies of other professional fields, which range from 0.75 in Meteorology to just 0.37 in Clinical Psychology. Coefficients in the 0.6 and above range are viewed very favorably. Different critics have different preferences, and so we should not expect perfect uniformity in assessments.

In that light, our results show that there is ample critical consensus between the four publications. Correlating the pairs of ratings, we find a high degree of consensus \( r = 0.64 \) between International Wine Cellar and each of Spectator and Advocate. We observe a lower, but still moderately high degree of consensus between Spectator and Advocate \( r = 0.47 \). Wine Enthusiast has low correlation with IWC and Spectator \( (0.17 \text{ and } 0.36, \text{ respectively}) \), but a moderately high correlation with Advocate \( r = 0.51 \).

An important question is whether the conditions under which the wines were tasted affects the level of consensus between the critics. Of the four, Spectator has the most stringent blind tasting policy, blind to both price and winemaker, Enthusiast tastes blind to price, and Advocate and IWC have no blind tasting policy. However, there is no clear connection between tasting policy and consensus. Spectator shows a higher degree of consensus with Advocate and IWC that with Enthusiast.

We find that these correlations vary when we split the samples by varietal, examining Merlot, Cabernet Sauvignon and Chardonnay separately. There is less concordance between
Spectator and each of Enthusiast and Advocate in Cabernet than the other varietals. Advocate and Enthusiast show high correlation on the red varietals but almost none on Chardonnay.

Further, we conduct several robustness checks. There is little difference when examining rank correlations, suggesting that the influence of outliers, and the arbitrariness of the cardinal rating scale, is small. We also conduct a regression analysis to determine if average differences in ratings per varietal or by state play a role. The pattern of correlation holds in this analysis as well.

RELATED LITERATURE

By far, most of the literature on expert wine rating consensus analyzes wine judge performance in major wine competitions. The well-known 1976 “Judgment in Paris” spawned several studies comparing the ratings of wine judges, reevaluating the results of the competition, and developing new methods of comparison (see, e.g., Ashenfelter and Quandt, 1999; Cicchetti, 2004a, 2004b, 2006). More recently, a tasting of New Jersey and French wines at the 6th annual meeting of this association, a competition suggesting no distinguishable difference in wine quality among these wines, garnered much attention (Ashenfelter and Storchmann, 2012; Bodington, 2012; Ginsburgh and Zang, 2012; Quandt, 2012; Taber, 2012; Ward, 2012).

In an early study, Brien, et al. (1987) found intra-judge reliability to be reasonably high based on tasting and re-tasting of the same wines, especially if the re-tasting occurred in the same day. Lawless et al, (1997) found a large range of reliability across wine judges, and also found that reliability of the mean ratings was higher than the mean reliability of the individual judges, a result that is consistent with the literature on judge reliability across many fields (Ashton, 2011). In contrast to these early studies, Cliff and King (1997)
suggested that wine judges’ ratings appear to be random. Gawel and Godden (2008) also reported a large variability in reliability across judges, with a mean reliability correlation coefficient of 0.45. Further, they find a mean absolute difference in scores of repeated tastings to be 1.04 for reds and 1.16 for whites on a 20-point scale, representing an average difference of about 5%.

Following Ashenfelter’s (2006) discussion of the general lack of concordance among wine judges, Hodgson (2008) analyzed wine judge performance, defined as a combination of consensus and reliability of judges. In his analysis of the 2005-2008 California State Fair Commercial Wine Competition, Hodgson found a high level of temporal inconsistency on ratings of the same wine by the same judge. Overall, he concluded that in only 46% of the cases was the wine quality alone the significant factor in determining a judge’s score. In a follow-on study, Hodgson (2009b) also concluded that few wine judges could be considered experts, in terms of reliability. Indeed, Hodgson concludes that fewer than 30% obtain that status.

Cao and Stokes (2010) went further, dissecting judge performance at a 2009 California wine competition into parameters for bias, discrimination and variability, which were estimated with a Bayesian ordinal model. Here bias refers to the difference between a judge’s scores and the mean scores among all judges of the same wine. Discrimination refers to the similarity between the ranking assigned by a judge and the rankings of the other judges, a measure of consensus. They concluded that only three out of sixty-seven judges at the wine competition displayed poor discrimination ability and excessive variation in their scores.

Hodgson (2009a) also extended his analysis to thirteen major wine competitions, where the competition was the observation, not the individual judge. The research questions were related to the replicability of a wine’s medal awards across competitions. Hodgson
found that among 375 wines entered in five competitions, no wine received a gold medal in either all five or four. Only 6 wines received three gold medals, 20 received two and 106 received only one gold medal. Furthermore, and perhaps most damaging for competition concordance, of the 106 wines receiving a gold medal in one competition, 84 of them received a “no award,” the lowest score, in another competition. Correlation results also suggest lack of concordance with a maximum inter-competition correlation coefficient 0.33. Hodgson concluded that receiving a gold medal can be statistically explained by chance alone.

The analysis of ratings of prominent wine critics, that which we undertake here, is less common than analysis of the performance of wine judges at competitions. As far as we are aware, only two studies have examined the degree of consensus among professional wine critics: Gokcekus and Nottebaum (2011) and Ashton (2013). As mentioned above, the former study focused on the relationship between prominent wine critic ratings and “community” ratings: the average of ratings given by anonymous, and amateur, subscribers to Cellar Tracker, a website dedicated to wine reviews. Gokcekus and Nottebaum found very high correlation between the professional ratings and the community ratings ($r = 0.77, 0.77, 0.83$). However, the community ratings may have been influenced by the professional ratings themselves – users of Cellar Tracker have direct access to the professional ratings. More recently, Ashton (2013) correlated ratings of prominent wine critics of red Bordeaux wine. He found that each of 100 pairwise correlations was statistically significant at the 0.01% level and that the mean of all correlations was 0.6. Notably, he confirmed a suspicion of many oenophiles that leading critics Robert Parker and Jancis Robinson have divergent preferences, reflected in an average ratings correlation coefficient (across seven vintages) of 0.45, which was the lowest of all pairs of critics examined.
In previous research we compiled data on wines from California and Washington State that had been reviewed by *Wine Spectator*. Two hundred each of Cabernet Sauvignon, Merlot and Chardonnay were selected at random from all California and Washington wines from 2005 reviewed by *Spectator*. Of the major varietals grown in California and Washington these are the most common and also the most likely to produce extraordinary wines.

### Table 1: Sample statistics, by Varietal

#### Panel A: Merlot

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine Spectator Rating</td>
<td>200</td>
<td>85.115</td>
<td>5.125733</td>
<td>68</td>
<td>93</td>
</tr>
<tr>
<td>Wine Advocate Rating</td>
<td>27</td>
<td>89</td>
<td>2.660249</td>
<td>82</td>
<td>94</td>
</tr>
<tr>
<td>International Wine Cellar Rating</td>
<td>28</td>
<td>88.17857</td>
<td>2.21198</td>
<td>85</td>
<td>92</td>
</tr>
<tr>
<td>Wine Enthusiast Rating</td>
<td>103</td>
<td>87.13592</td>
<td>3.304924</td>
<td>81</td>
<td>95</td>
</tr>
<tr>
<td>Release price</td>
<td>200</td>
<td>32.225</td>
<td>20.54421</td>
<td>8</td>
<td>140</td>
</tr>
</tbody>
</table>

#### Panel B: Cabernet Sauvignon

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine Spectator Rating</td>
<td>200</td>
<td>87.865</td>
<td>3.981966</td>
<td>75</td>
<td>97</td>
</tr>
<tr>
<td>Wine Advocate Rating</td>
<td>34</td>
<td>91.38235</td>
<td>3.025294</td>
<td>86</td>
<td>100</td>
</tr>
<tr>
<td>International Wine Cellar Rating</td>
<td>56</td>
<td>89.96429</td>
<td>2.885476</td>
<td>85</td>
<td>96</td>
</tr>
<tr>
<td>Wine Enthusiast Rating</td>
<td>87</td>
<td>89.31034</td>
<td>3.606552</td>
<td>81</td>
<td>97</td>
</tr>
<tr>
<td>Release price</td>
<td>200</td>
<td>68.75</td>
<td>51.84774</td>
<td>8</td>
<td>300</td>
</tr>
</tbody>
</table>

#### Panel C: Chardonnay

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine Spectator Rating</td>
<td>200</td>
<td>87.12</td>
<td>3.413776</td>
<td>77</td>
<td>95</td>
</tr>
<tr>
<td>Wine Advocate Rating</td>
<td>51</td>
<td>91.54902</td>
<td>2.759085</td>
<td>84</td>
<td>99</td>
</tr>
<tr>
<td>International Wine Cellar Rating</td>
<td>59</td>
<td>89.50847</td>
<td>2.654683</td>
<td>85</td>
<td>96</td>
</tr>
<tr>
<td>Wine Enthusiast Rating</td>
<td>120</td>
<td>87.95833</td>
<td>3.640103</td>
<td>81</td>
<td>96</td>
</tr>
<tr>
<td>Release price</td>
<td>200</td>
<td>31.98</td>
<td>16.93439</td>
<td>8</td>
<td>88</td>
</tr>
</tbody>
</table>

From Table 1, we can see that the prices range from $8 to $300, with the mean price substantially higher for Cabernet than the other two varietals. The mean expert ratings vary from 85.1 for *WS* and Merlot to 91.5 for *Wine Advocate* and Chardonnay. For all varietals,
*Advocate* had the highest average ratings. This is likely due to differences between how *Advocate* and the others select wines to review. *Wine Advocate* is focused on reviewing and recommending a narrow selection of excellent wines, while other publications such as *Spectator* and *Enthusiast* review a wider range of wines. In the case of *Spectator*, wines reviewed include those submitted by winemakers as well as wines chosen by *Spectator* staff. Interestingly *International Wine Cellar* exhibits a lower average rating than *Wine Advocate*, though its scope is similarly narrow, which reflects noted “grade inflation” *Wine Advocate* ratings that *IWC* has resisted.

These publications were selected because they represent the critical opinions that are most often consulted for the U.S. market. Winemakers and retailers frequently cite their ratings in marketing materials. Further, the four have adopted nearly identical rating scales, in which 100 represents perfection, and most wines are categorized in the 75 – 99 point range. The 100 point scale seems to be the standard in America, while European and Australian critics tend to go with the 20 point scale. Perhaps this is due to an American conflation of 100% with perfection. For each publication, the 100 point scale is broken down into quality categories, typically 100-95, 94-90, 89-85, etc (*Advocate* uses 100-96 as its top category).

*The Wine Advocate* was started by prominent critic Robert Parker Jr. in 1978. In our 2005 sample of Washington and California wines, Robert Parker and his associate Antonio Galloni tasted and reviewed California wines, and another associate, Pierre Rovani, covered Washington wines. *Advocate* has no blind tasting policy, allowing its tasters full knowledge of prices, winemakers, region of origin and varietal. However, it eschews the receipt of compensation, gifts and favors from winemakers and industry representatives.¹

¹ See [www.erobertparker.com](http://www.erobertparker.com) for tasting philosophy.
Spectator has been reviewing wines since 1976. It has an explicit tasting policy that requires all tasters to be blind to price and winemaker. In our sample, James Laube was the primary taster for California Cabernet Sauvignon and Chardonnay, though MaryAnn Worobiec reviewed some. Tim Fish reviewed California Merlot for Spectator, and Harvey Steiman reviewed all wines from Washington. Spectator reports that 25 years of blind tastings have produced an approximately normal distribution of wine ratings, centered on 86, with only 21% classified as “outstanding” (90-94) and just 2% classified as “classic” (95-100).²

Enthusiast reports tasting blind to price, but not winemaker, since 1999.³ Critics for the 2005 vintage were Steve Heimhoff, for California wines, and Paul Gregutt for Washington. Information about critics for International Wine Cellar was not available - Stephen Tanzer, who founded it, is its primary critic, but it is not clear whether any associates were the tasters for wines in our sample. Also, no information on tasting policy could be found on IWC’s official site.⁴ Table 2 summarizes the critics and tasting policy of each publication.

---


³ www.winemag.com

⁴ http://www.wineaccess.com/expert/tanzer/newhome.html
We review the correlations between rating scores for each pair of publications, which are presented in Table 3. The first column of correlations contains the correlations with all wines included, while the following three columns look at each varietal separately.

From Table 3, we can see that while a high degree of correlation exists between some publications, it is not uniform. In the case of \textit{WE} – \textit{IWC}, there is no statistically significant correlation. Ashton (2013), (having reviewed the literature on critical consensus in many
fields) characterizes a level of consensus around 0.60 as “quite favorable”. This represents a higher level of consensus than of ratings of wine professionals who are not prominent critics, as found in his survey of earlier studies (Ashton, 2012). Here we find levels of consensus above 0.60 between Spectator and IWC, as well as between Advocate and IWC. The 0.60 threshold is met between Spectator and Advocate, but only for Merlot. Overall, they display a moderate degree of consensus, with a correlation of 0.47.

Enthusiast displays a lower degree of consensus with the other publications. Its overall correlation with IWC is statistically zero, and just 0.36 with Spectator. Advocate is more highly correlated with Enthusiast, at 0.51, which evidently comes from a combination of high correlation in Merlot and Cabernet and zero correlation with Chardonnay.

It would be a provocative finding if consensus were linked to tasting policy, if blind reviewers were more in accord than non-blind reviewers. However, there is no clear pattern from Table 3 regarding tasting policy and consensus. Spectator, with the strictest tasting policy, has its least consensus with Enthusiast, which has the second-most strict tasting policy.

RANK CORRELATIONS

Given that wine scores assign a precise cardinal value to what is essentially an ordinal preference ranking, we go further by examining the degree of agreement in ordinal rankings between the four publications. In Table 4, we present the correlations in rankings.

These values are generally similar to those of the (cardinal) correlations presented in Table 3, but some differences emerge. The rank correlations between Spectator and Advocate, and Enthusiast and Advocate, are somewhat lower. This may reflect overlap in the high- and low-ends of the distribution of scores that covers up a lower degree of concordance.
in the middle. The rank correlation between Spectator and Enthusiast is somewhat better than the general correlation, but still low at 0.39. The concordance between Enthusiast and IWC is still statistically nil. Overall, the rank correlations support what we find with the correlations of Table 3.

Table 4: Rank correlations (Spearman’s rho)

<table>
<thead>
<tr>
<th>Pair</th>
<th>All varietals</th>
<th>Merlot</th>
<th>Cab. Sauv.</th>
<th>Chard.</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS - WE</td>
<td>0.39 (310)</td>
<td>0.43 (103)</td>
<td>0.24 (87)</td>
<td>0.35 (120)</td>
</tr>
<tr>
<td>WS - WA</td>
<td>0.41 (112)</td>
<td>0.54 (27)</td>
<td>0.34 (34)</td>
<td>0.47 (51)</td>
</tr>
<tr>
<td>WS - IWC</td>
<td>0.67 (143)</td>
<td>0.63 (28)</td>
<td>0.68 (56)</td>
<td>0.62 (59)</td>
</tr>
<tr>
<td>WA - IWC</td>
<td>0.66 (35)</td>
<td>0.74† (7)</td>
<td>0.80† (10)</td>
<td>0.65 (18)</td>
</tr>
<tr>
<td>WE - WA</td>
<td>0.44 (62)</td>
<td>0.70 (17)</td>
<td>0.68 (17)</td>
<td>0.13† (28)</td>
</tr>
<tr>
<td>WE - IWC</td>
<td>0.18† (86)</td>
<td>0.28† (18)</td>
<td>-0.08† (31)</td>
<td>0.24† (37)</td>
</tr>
</tbody>
</table>

Note: Number of observations in parentheses.
† Statistically insignificant at the 5% level.

REGRESSION ANALYSIS

We conducted a regression analysis to examine whether the linear relationships given by the correlations presented in Table 3 held up after controlling for average differences in the ratings by varietal and state. Each regression estimates an equation of the form:

\[
\text{rating}_{1i} = \alpha + \beta_1 \text{rating}_{2i} + \beta_2 \text{Cabernet}_i + \beta_3 \text{Merlot}_i + \beta_4 \text{Wash}_i + \epsilon_i \tag{1}
\]

where \(i = 1, \ldots, n\) indexes wines and \(\epsilon\) is the random disturbance term.

Each column of Table 5 shows the results of OLS regression of one publication’s rating on another, with controls for Cabernet Sauvingon, Merlot, and Washington State. In each of the first three, with WS rating as the dependent variable, the Washington State indicator is positive and statistically significant, indicating that the average WS rating is higher for wines from Washington than wines from California. In regressions (3) and (5) the

---

\(^5\) On its face the lack of relationship between WE and IWC may seem at odds with both of them being related to the other publications’ ratings. This may be due to a feature of our sample: the set of wines in each pair-wise sample differs.
Merlot indicator is significantly negative, showing that WS and WA tend to rate Merlot wines lower than those of Cabernet Sauvignon and Chardonnay. In terms of the rating slope coefficients, the control variables did not matter for the relationship between ratings: pairs of ratings that were more highly correlated in Table 3 have larger slopes in Table 5.

We conducted tests for the hypothesis that $\beta_1 = 1$, and only for regression (4) could we not reject that hypothesis. That is, the relationship between *The Wine Advocate* and *International Wine Cellar* is so close as to be in lockstep, as every extra point Stephen Tanzer (or an associate) assigns to a wine can be expected to be matched with a point by Robert Parker. This high degree of concordance is not observed for the other pairs of critics.

In Table 6 we examine the average differences within each publication due to varietal and Washington State. This is similar to the approach in Table 5, but we omit the rating independent variable to simply examine how the average ratings vary. That is, we estimate:

$$ \text{rating}_i = \alpha + \beta_1\text{Cabernet}_i + \beta_2\text{Merlot}_i + \beta_3\text{Wash}_i + \epsilon_i $$  \hspace{1cm} (2)

Under the assumption that $(\epsilon_i | \text{Cabernet}_i, \text{Merlot}_i, \text{Wash}_i) = 0$,

$$ E(\text{rating}_i) = \alpha + \beta_1\text{Cabernet}_i + \beta_2\text{Merlot}_i + \beta_3\text{Wash}_i $$  \hspace{1cm} (3)

so the interpretation of the estimated coefficients of *Cabernet*, *Merlot* and *Wash* is the difference in the average rating for wines of those categories.
We see in Table 6, column (1), that there is a large gap (3.8 points) in the average Wine Spectator ratings for Washington wines over California wines. It cannot be determined with our data whether this gap is due to more generosity on the part of Spectator’s Washington critic, Harvey Steiman, or whether it is simply a reflection of strong quality of Washington wines, since we don’t have a second rating from any of the other Spectator critics for comparison. If all of this 3.8 point gap were due to scaling differences between Spectator critics and not true quality differences, this could help explain the persistent price gaps between Washington and California wines found in recent research by Miller and Stuen (2013). Despite this large difference in average ratings, the correlation coefficients of ratings among Washington wines alone are very similar to those reported in Table 3 with California wines included. Within the set of Washington wines with pairs of ratings available, the WS-WA correlation is 0.43 and the WS-IWC correlation is 0.58. These are lower than the overall correlations, but still in the same range of consensus.

A similar but smaller Washington point gap is observed for Enthusiast, as seen in column (3). The average ratings for IWC do not vary by varietal and state, as seen in column (4), but Advocate discounts Merlot by about 2.3 points on average, shown in column (2).

<table>
<thead>
<tr>
<th>Dep. Var.:</th>
<th>WS rating</th>
<th>WA rating</th>
<th>WE rating</th>
<th>IWC rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>=1 if Cabernet</td>
<td>0.357</td>
<td>-0.062</td>
<td>1.135*</td>
<td>0.437</td>
</tr>
<tr>
<td>=1 if Merlot</td>
<td>-3.029</td>
<td>-2.264**</td>
<td>-1.382**</td>
<td>-1.378</td>
</tr>
<tr>
<td>=1 if Washington</td>
<td>3.822**</td>
<td>-0.536</td>
<td>2.093**</td>
<td>0.820</td>
</tr>
<tr>
<td>Intercept</td>
<td>87.032**</td>
<td>91.602**</td>
<td>87.766**</td>
<td>89.439*</td>
</tr>
<tr>
<td>R-sq</td>
<td>0.164</td>
<td>0.133</td>
<td>0.107</td>
<td>0.065</td>
</tr>
<tr>
<td>Obs.</td>
<td>401</td>
<td>112</td>
<td>310</td>
<td>143</td>
</tr>
</tbody>
</table>

* Significant at 5% level.  ** Significant at 1% level.
CONCLUDING COMMENTS

We undertook a straightforward analysis of the degree of consensus among prominent critical publications in the U.S. The degree of consensus, as measured by the correlation coefficient of wine quality ratings, varied widely between pairs of critics and also by varietal. Among these publications, *Wine Enthusiast’s* opinion diverges from the others the most. Excluding *Enthusiast*, we conclude that the level of consensus in wine ratings by professional critics in the U.S. market is high, and similar to the levels Ashton (2013) found for consensus among critics of Bordeaux wine. Further, the level of consensus between each pair of *Spectator, Advocate* and *IWC* substantially exceeds the level of consensus between wine competition judges, as Ashton (2012) reports the mean correlation coefficient across many studies to be just 0.34 (from, *inter alia*, Brien, May and Mayo, 1987; Cicchetti, 2004a; Hodgson, 2009a; Ashton, 2011).

An optimistic and conventional explanation for the greater consensus between critics than judges is that it is due to more extensive experience evaluating and comparing wines. It is possible though that critics are influenced by knowledge of price, winemaker and possibly also the ratings of other critics, which could lead to greater similarity of ratings. However, we do not observe greater consensus among non-blind critics than with the more heavily blinded critic *Spectator*.

It would be useful to expand this study to include a wider set of wines. We initially sampled six hundred wines, but only found one or more matched ratings for 401 of them. For most pairs the correlations were statistically significant, but for some pairs we did not have enough observations within each varietal. More data collection and coverage of more wine varietals and vintages would serve to buttress our results.
References


16


