Journal of Wine Economics, Volume 12, Number 4, 2017, Pages 395–404
doi:10.1017/jwe.2017.38

Wine Competitions: Reevaluating the Gold Standard*

Christopher Bitter [a]

Abstract

Competition medals are one of the most readily available sources of expert opinion to wine consumers, yet the "expertise" of competition judges and efficacy of medals have been questioned in the literature. This paper reevaluates the relevance of gold medals using data from ten competitions and scores from two leading wine publications. The analysis begins by exploring differences in gold medal award rates across competitions while holding wine quality constant through paired comparisons, which are found to be substantial. Next, the relevance of gold medals as indicators of wine quality is assessed, using the average scores from *Wine Enthusiast* and *Wine Spectator* as surrogates for quality. By itself, knowledge that a wine is a gold medal winner appears to have little relevance, as these wines do not score significantly higher than other medal winners. However, evidence suggests that golds from some competitions may be more relevant than others. (JEL Classifications: L15, L66)

Keywords: wine competitions, wine judging agreement, wine quality evaluation.

I. Introduction

Wine is an experience good, meaning its quality cannot be assessed prior to its consumption. When choosing an experience good, consumers must rely on external clues, such as price, reputation, or opinions of others who have already tried the product. In the case of wine, the opinions of "experts" play a particularly important role. The medals conferred by wine competitions are one of the most widely available sources of expert opinion, although the "expertise" of the typical competition judge has been questioned in the literature.

Despite this fact, wine competitions continue to proliferate in number. A cursory examination reveals nearly one hundred active competitions in the United States alone, with new ones popping up almost every year.

* I would like to thank the anonymous referee, the editor (Karl Storchmann), and the participants of the 11th Annual AAWE Conference for their helpful suggestions that improved the paper.
[a] Vintage Economics, 1506 N. 80th St., Seattle, WA 98103; e-mail: Bitter@VinEconomics.com.

This paper investigates the relevance of gold medals to the consumer. To be relevant, a gold medal must be useful in differentiating higher-quality wines from those of lower quality, as medal awards are typically unaccompanied by other information that would aid purchasing decisions. Moreover, gold medals should represent a consistent standard of quality across competitions and over time, as their meaning would be diminished if some competitions were more generous with their awards than others.

Section II provides a brief review of the literature, followed by a description of the data used in the analysis. Sections III.A and III.B examine differences in the prevalence of gold medal awards across competitions, and Section III.C assesses the relevance of gold medals as indicators of wine quality via the average scores from two prominent wine publications, as well as the extent to which the gold-medal standard has inflated over time. Section IV concludes.

II. Literature

Much has been written about the inadequacy of expert opinion, particularly regarding wine competitions. It is impossible to precisely measure accuracy in the case of wine-quality evaluations, because no universally agreed-upon external criterion exists. However, for expert judgments to be accurate and objective measures of quality, two potentially observable conditions must be met. First, judges must exhibit reliability, meaning that they can replicate their own findings in subsequent evaluations of the same wine; second, judges must agree with one another in their evaluations, a metric known as *consensus* (Ashton, 2012).

Prior research has found competition judges to be lacking when it comes to both metrics. Based on a review of prior studies, Ashton (2012) finds a mean correlation between judges' own scores in repeated tastings of the same wine of just 0.5. However, reliability varies widely across judges, as some exhibit far greater consistency in scoring than others. In other words, some "experts" are more "expert" than others.

Hodgson (2008) provides a poignant example of the lack of reliability in a study focused on the California State Fair wine competition, where each judge was served four identical wines three times within a flight of thirty wines. Judges were able to assign the replicates to the same medal category just 18 percent of the time, and this repetition was generally for wines of the lowest quality that were not awarded medals. Only 10 percent of judges assigned all four replicates to the same medal category, while another 10 percent awarded at least one of the four a gold medal in one trial and a bronze medal or no award in another.

This lack of reliability should not be entirely surprising—after all, competition judges are humans, not machines, and the task they face is daunting, as they are often required to taste dozens of wines over a short period of time that range

widely in quality and style. The position of a wine in the lineup is also a wild card, as the sensory attributes of prior wines may influence perceptions of subsequent wines. The fact that some judges perform better than others should also not be a surprise, as they vary in terms of level of experience, ability, and knowledge.

Wine-competition judges fare even worse when it comes to the consensus metric, as Ashton (2012) reports a mean correlation in scoring across judges of only 0.34. Why can judges not reach greater consensus? The lack of reliability is clearly part of the explanation, as it introduces a random component into the scores. But there is more to it than that. Positive correlations indicate that judges share at least some criteria relating to wine quality, but they, like consumers, have differing biological makeups that influence their perceptions of tastes and smells, and their prior experiences with wine vary. Thus, it is unreasonable to expect their evaluations to be entirely objective.

Moreover, Cao and Stokes (2010) show that the lack of consensus among judges is due in part to the fact that some systematically score wines higher or lower than the average, which is referred to as *bias*, and that some use narrower scoring ranges than others—in other words, they "discriminate" less between "good" and "bad" wines. Thus, the scoring scale itself is subjective.

The lack of consensus and reliability can translate to arbitrariness in medal awards, as demonstrated by Hodgson (2009). He examines wines entered in each of five different competitions and finds that 98 percent of those that won at least one gold did not receive an award, or was awarded a bronze medal, in at least one other competition. Overall, there was little correlation (0.11) between awards across competitions. The judgments were most consistent for wines rated as average or below average, which leads Hodgson (2009, p. 5) to conclude that "wine judges concur in what they do not like but are uncertain about what they do."

If gold medal awards are truly arbitrary, they have little relevance to the consumer. However, prior research has not attempted to directly relate competition awards to an external measure of wine quality or explore the possibility that some competitions may be more proficient than others. The finding that reliability varies greatly among judges implies the possibility that gold medals from competitions employing skilled judges and more rigorous tasting formats could have greater relevance.

III. Empirical Analysis

A. Data

The data used in the empirical analysis include medal awards for Washington State wines from ten competitions, obtained directly from competition websites as well as greatnorthwestwine.com, which publishes results from a number of competitions. These data include only entries that won awards, as none of the competitions provides lists of losers, and they cover the period from 2013 to 2017, although data is

unavailable for the first or last year in several cases. The competitions, shown in Table 1, are chosen primarily on the basis of the availability and format of the award data and the number of observations.

The empirical analysis also uses scores for nearly three thousand Washington wines reviewed by *Wine Enthusiast* and *Wine Spectator* between 2012 and 2016.

To facilitate comparisons across the twelve sources, I first reconstruct wine names using a consistent format, as naming conventions vary. I then match names across sources and verify them to the extent possible using additional information, such as prices. I believe the matches to be reasonably accurate, although it is inevitable that a few false positives have not been detected.

B. The Gold Standard

The first question I address is whether a consistent standard of selectivity exists in gold medal awards across competitions. Based on an informal analysis of published medal awards and unofficial entry totals from more than a dozen national and regional competitions over the last several years, gold medal award rates (including double golds) vary widely. They typically range from 15 to 25 percent, but they exceed 40 percent at the Seattle Wine Awards (SWA) and are less than 10 percent at TexSom. These differences could be attributable to variation in the quality of entries or to differences in selectivity.

To sort this out, I analyze the set of overlapping medal winners between pairs of competitions, which ensures strict comparability in terms of quality of entrants. I use the SWA as the reference competition against which to compare the other nine because it generates the greatest number of matched observations.

The results, shown in Table 2, demonstrate that substantial differences in selectivity exist across competitions. Each column represents a comparison between the competition named at the top (the subject) and the SWA. The top row indicates the number of wines that received medals in both competitions, and the second and third rows give the proportion of these wines that were awarded gold medals (including double golds) by the SWA and the subject competition, respectively. The final row is simply the ratio of rows two and three. The SWA is the most generous competition by far, as medal winners are awarded golds at more than twice the rate of any other. Conversely, TexSom appears to be the most selective—it awards golds at just one-sixth the rate of the SWA. Thus, a gold medal does not imply a consistent standard of selectivity.

C. Is a Gold Medal a Relevant Indicator of Quality?

For gold medals to be relevant to consumers they must also be able to distinguish the high-quality entries from the pack. This proposition is challenging to test due to the

Table 1
Wine Competitions Included in the Analysis

Competition	*Observations*	*Type*
Dan Berger's International Wine Competition	352	International
Cascadia Wine Competition	2366	Regional
Great Northwest Invitational Wine Competition	971	Regional
Pacific Rim International Wine Competition	227	International
Savor Northwest Wine Awards	491	Regional
Seattle Wine and Food Experience	240	Regional
San Francisco Chronicle Wine Competition	1456	National
San Francisco International Wine Competition	856	International
Seattle Wine Awards	4438	Regional
TexSom International Wine Awards	459	International

lack of an external wine-quality criterion. However, it has been shown that professional wine critics exhibit greater consensus than do competition judges (Ashton, 2012, 2013). This trait is likely due to their superior skills as well as to the settings in which they taste the wines. Moreover, Ashton (2011) demonstrates that a "composite" judgment based on the average score of multiple critics is generally more accurate than that of any of the individual judgments upon which it is based, and that most of the improvement can be achieved by considering the scores of only two or three judges.

Based on this reasoning, I use the average score for the set of wines reviewed by *Wine Spectator* and *Wine Enthusiast* as a surrogate for quality (hereafter referred to as the "critic's score"). I choose these publications because they have the greatest overlap with the competition entries, and both employed experienced and reputable wine writers throughout the study period. The correlation between their scores is 0.42, which appears to be fairly typical for professional wine reviewers (Ashton, 2012). It would have been desirable to include more publications, but doing so would have severely limited the number of observations available for analysis. Although my approach is clearly imperfect, it should be sufficient to draw useful insights regarding the connection between gold medals and wine quality.

In some cases, critics' scores are published prior to a wine's appearance in a competition, and in others they appear after. However, based on content from their respective websites, I believe that both publications and all ten competitions taste blind, so their judgments should be independent.

The data in Table 3 pertain to the 849 wines that medaled in at least one of the ten competitions over the period from 2014 to 2016 and have scores from both publications. Approximately 39 percent medaled in only one competition, 26 percent received medals in two, and 35 percent medaled in three or more. Because the competitions only publish lists of winners, it is impossible to identify the number of competitions each wine was entered in. Sixty percent of the wines received at least one

Table 2
Gold Medal Awards: Paired Comparisons

Metric	*Dan Berger*	*Cascadia*	*Great Northwest*	*Pacific Rim*	San Fran. Chronicle	*San Fran. Intl.*	*Savor Northwest*	*Seattle Wine&Food*	*TexSom*
Observations	126	679	362	94	342	207	189	84	149
SWA golds	62.7%	63.5%	66.9%	70.2%	58.5%	67.6%	68.3%	59.5%	63.8%
Subject golds	18.3%	20.3%	24.6%	24.5%	26.0%	20.8%	28.6%	28.6%	10.7%
Gold ratio	3.4	3.1	2.7	2.9	2.2	3.3	2.4	2.1	5.9

Table 3
Critics Score by Medal Award

	Highest Medal Awarded			*Number of Golds Awarded*			
Metric	*Bronze*	*Silver*	*Gold*	*0*	*1*	*2*	*3+*
Count	89	252	508	341	365	106	37
Percentage of total	10.5%	29.7%	59.8%	40.2%	43.0%	12.5%	4.4%
Average score	88.6	88.7	88.9	88.7	88.9	89.0	89.2
Percentage of 90+	27.0%	29.4%	33.5%	28.7%	32.9%	34.9%	35.1%

gold medal. It is also likely that many wines won golds in competitions not included in the database.

The first panel depicts the average critics' scores and percentages of wines achieving scores above 90 points, segmented by the highest medal awarded to each wine. The results suggest that consumers have little to gain from buying gold medalists, as the average score and proportion receiving 90+ points are only slightly higher than those of bronze or silver medalists. The second panel shows the average critics' scores based on the number of gold medals each wine won. Again, there is only a slight improvement with each additional gold, so this knowledge would not be particularly relevant either.

The aggregate results presented above do not necessarily imply that gold medals are entirely irrelevant. Indeed, because the quality of judges, tasting procedures, and selectivity vary across competitions, some may be more relevant than others. To test this proposition, I examine differences in the mean critics' scores for wines that received gold medals versus those awarded bronze or silver medals for each competition.

As indicated in Table 4, the difference in means ranges from –0.3 to 1.1 points but based on *t*-tests is only statistically significant in three cases. This result implies that some competitions are better at differentiating between low- and high-quality wines than others, which is likely attributable to the factors alluded to above. For example, gold medal winners scored a full point higher in the TexSom competition, and the mean critics' score for golds exceeded 90 points. More than half of the 2016 TexSom judging panel had Master of Wine or Master Sommelier designations, and the evaluation methods articulated on its website are far more detailed and rigorous than is typical. A 1-point difference may not have great practical importance to the consumer, but selecting a gold medal winner from either of the competitions at the top of the list does appear to improve the odds of obtaining a high-quality wine.

Finally, I estimate a set of simple binomial logit models for the six competitions with sufficient observations to evaluate whether the gold standard has inflated over time. The dependent variable is the probability of receiving a gold medal, and the determinants are time (entry year) and quality (critics' scores). The latter

Table 4
Competition Comparison: Mean Critics' Scores by Award Type

	Count		*Percentage*	*Average Score*			*Percentage Scoring 90+*		
Competition	*B/S*	*Gold*	*Gold*	*B/S*	*Gold*	*Diff*	*B/S*	*Gold*	*Diff*
Dan Berger	63	16	20%	88.5	89.5	1.1*	22.2%	43.8%	21.5%
TexSom	203	24	11%	89.1	90.1	1.0**	35.5%	58.3%	22.9%
Savor Northwest	66	25	27%	88.8	89.3	0.5	27.3%	36.0%	8.7%
Cascadia	283	98	26%	88.6	89.1	0.5*	26.5%	33.7%	7.2%
San Fran. International	136	35	20%	88.3	88.7	0.4	18.4%	25.7%	7.3%
San Francisco Chronicle	257	104	29%	88.7	88.8	0.1	27.2%	28.8%	1.6%
Seattle Wine & Food	42	18	30%	88.3	88.4	0.1	23.8%	33.3%	9.5%
Seattle Wine Awards	194	408	68%	88.9	88.9	0.1	33.5%	34.1%	0.6%
Great Northwest	209	71	25%	89.4	89.1	0.3	42.6%	35.2%	−7.4%
Pacific Rim	29	13	31%	88.7	88.4	0.3	27.6%	7.7%	19.9%

Note: Statistical significance levels are * (5%) and ** (1%).

Table 5
Logit Model Results

	Cascadia	*Great NW*	San Francisco Chronicle	*San Francisco International*	*Seattle Wine Awards*	*TexSom*
Intercept	−16.09**	5.57	−3.63	−13.58	−0.81	−34.32**
Quality	0.16*	−0.08	0.02	0.13	0.02	0.37**
Time	0.27**	0.15	0.17*	0.15	−0.01	−0.38
N	381	280	361	171	602	227
Pseudo-R^2	0.03	0.01	0.01	0.02	0.00	0.06

Note: Statistical significance levels are * (5%) and ** (1%).

variable is used to control for differences in the quality of entrants over time, which could influence gold medal award rates.

The results are shown in Table 5. Neither quality nor time has a substantial impact on the odds of a wine winning a gold medal. Quality is statistically significant in just two cases, which is consistent with the results discussed in the prior section. The time coefficients are positive in four models but significant in only two: Cascadia and *San Francisco Chronicle*. Thus, the results do not indicate widespread inflation in the gold standard, at least for this small set of competitions. It is possible, however, that the critics have become more generous in their scoring, which would mask grade inflation in the competitions.

IV. Conclusion

The results of the analysis largely confirm those of prior work but also generate several novel and potentially important insights.

First, there are tremendous differences in generosity across competitions—some are much more selective when it comes to awarding golds than are others—but there does not appear to be widespread inflation in the gold standard. Second, knowledge that a wine received a gold medal, by itself, does not appear to be particularly relevant to the consumer, as in the aggregate, wines receiving gold medals do not achieve significantly higher critics' scores than those receiving bronze or silver medals. The final, and perhaps most intriguing, finding is that all golds are not created equal—some competitions appear to be able to more effectively differentiate between low- and high-quality wines than others. However, from a practical standpoint, it may be difficult for the consumer to identify these competitions.

Several limitations should also be noted. The analysis includes only a small subset of competitions, which limits the ability to generalize the results. The lack of

information on wines that did not earn medals is also a shortcoming, and it is possible that competitions are more effective at weeding out low-quality wines. Most importantly, the reliability of the findings depends on the efficacy of the critics' scores that are used as a surrogate for quality with which to judge the competitions.

References

Ashton, R. H. (2011). Improving experts' wine quality judgments: Two heads are better than one. *Journal of Wine Economics*, 6(2), 160–178.

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.

Cao, J., and Stokes, L. (2010). Evaluation of wine judge performance through three characteristics: Bias, discrimination, and variation. *Journal of Wine Economics*, 5(1), 132–142.

Hodgson, R. T. (2008). How expert are "expert" wine judges? *Journal of Wine Economics*, 4(2), 233–241.

Hodgson, R. T. (2009). An analysis of the concordance among 13 U.S. wine competitions. *Journal of Wine Economics*, 4(1), 1–9.