Evaluating Wine-Judging Events

Stephen Franzoi

A Suggested WVA Mantra

"Knowledge is power." This oft-quoted phrase can be traced back to the English philosopher Francis Bacon, an early advocate of the scientific method. Bacon's quote could easily be the "Winemakers' Mantra" for the Wisconsin Vintners Association. Seeking knowledge to increase our understanding of wine is why so many of us attend winemaking seminars, why our association regularly schedules wine education programs at monthly meetings, and why our members freely share new information about “all-things-wine.” This appreciation—and even love—of wine knowledge also means that we should never be afraid of new information, even when the accompanying knowledge challenges longstanding beliefs. We also should never try to protect others from nettlesome wine-related knowledge out of concern that they may not be able to put this information in proper perspective. With this philosophy in mind, I would like to present the following article on research that has evaluated wine judging, a “hot button” topic in some winemaking circles.

After making wine for a few years, many of us entered our creations into wine judging events. If you are like me, the first time you dipped your winemaking toes tentatively into the seemingly daunting waters of such competitions you were simply delighted by not being frost bitten by the judges’ evaluations. “A third place? Wow! That’s great!! Participation Ribbon? Fine and Dandy!! A Breathing Award for my wine? Ok, I’ll accept that too!” Such affirmations undoubtedly boosted your confidence and motivated you to further enhance your skills in the winemakers’ equivalence of the backstroke or butterfly. Perhaps you even began free styling in the ever-warming waters of all-that-is-winemaking. Yet it is also likely that the thrill you experienced over these early successes were occasionally replaced by dismay and discouragement when some of your winemaking entries failed to earn any awards at all. Maybe your entries even received an affirmative thumbs-down by the panel of judges who sniffed, tasted, and eyeballed them. Of course, negative feedback can be very instructive in perfecting skills in any area of pursuit. But what if you begin to suspect that the feedback is considerably less than perfect?
In many instances, this questioning of judges’ results is purely a case of "sour grapes" by those of us who have difficulty graciously accepting constructive criticism. Yet it is also true that judges are not infallible, nor should we expect them to be. Wine is a complex chemical beverage. Chemists have identified at least 400 aroma compounds that work on their own and with others to create complex flavors, with some noticeable at the first sniff and others detected only as an aftertaste. Further, perceptions of aroma and bouquet are based on far more than the chemicals found in wine; evaluating this complex liquid is indeed a challenge for anyone, and that challenge is made even more difficult when judges are asked to quantify their evaluations using a multiple-point scoring system. When it comes to evaluating a wine using the nose and the mouth without the aid of pH meters, acid testers, and sulfite meters, all judgments are, by definition, subjective, not objective. Unfortunately, many people’s understanding of the wine-judging process is muddled by the belief that when judges evaluate a wine they somehow can divorce themselves from their personal feelings, tastes, or unique life experiences, and also are not influenced by factors in the situation itself. Using a Star Trek analogy here to crystallize this misperception, far too many people—including many judges—believe that judges should be able to evaluate wines like the objective Vulcan Spock, when, in fact, they cannot escape their all-too-human Captain Kirk subjectivity. Noted wine expert Jancis Robinson concurs with this assessment, having declared on numerous occasions that wine tasting is so dependent on individual sensory equipment that it is decidedly subjective in nature, not objective.

If wine evaluation is indeed a subjective experience that is susceptible to human error, then how error ridden are wine-judging competitions? To answer this question, let’s examine research studies that focused on judges’ consistency in evaluating wines and also whether judges are influenced by the order in which they taste wines.

**How Consistent Are Wine Judges?**

In 2000 California winemaker and retired statistics professor Robert Hodgson became curious when some wines that he entered into multiple competitions earned very different evaluations. For example, his 1993 Zinfandel won the San Francisco International Wine Competition Best of Show yet earned no award in a subsequent event. Such experiences prompted Hodgson to set out on a systematic exploration to determine whether the inconsistencies in wine judging that he encountered were infrequent or commonplace in the world of wine competitions. Toward that end, in 2005, as a member of the advisory board for the California State Fair Commercial Wine Competition—which is the oldest contest of its kind in North America—Hodgson asked and received permission to conduct tests to check the reliability of the judges. In these tests, four-judge panels were presented with their usual flights of wine samples to sniff, taste, and examine. However, to measure the judges’ consistency, some of the wines were presented to the panel three times, and poured from the same bottle each time. Hodgson conducted these tests at the state fair each year over the next eight years. The wine evaluation scale used by the judges ranged from a low score of 50 to a high score of 100; most of the rated wines were given scores in the 70s, 80s, and low 90s.
Who were these judges? They were a *Who’s Who* of experts in the American wine industry, including professional winemakers, certified sommeliers, well-known wine critics and wine consultants, and even university professors who taught classes in winemaking and conducted research in winemaking. In this study, the judges would first indicate the wine’s score independently, which was then recorded. Afterwards, the judges discussed the wine, and based on the discussion, some judges modified their initial score. However, only the first independent score was used by Hodgson to analyze an individual judge’s consistency in scoring wines.

Results from the first four years of Hodgson’s study were published in 2008 in the *Journal of Wine Economics*, and they indicated that, overall, these wine experts were not very consistent judges. Over the three blind tastings, a typical judge’s scores for the “planted” wine varied by about plus or minus four points. For example, a wine rated as a good 90 would be rated a few minutes later by the same judge as an acceptable 86 and then a bit later as an excellent 94. Only about 10% of the judges were consistent in rating the identical wines, meaning that they gave the same wine presented to them three different times scores that varied by just plus or minus two points. In other words, these were the only judges whose ratings of the same wine typically stayed within the range of one medal. Yet even those judges who were consistent in their evaluations at Year 1 often were not very consistent at Year 2, so inconsistency was a more accurate descriptor of the vast majority of the judges’ performances. At the opposite end, another 10% of the judges gave the same wine far different ratings, ranging from a wine receiving a gold medal to the same wine receiving no medal at all.

In additional studies beyond California’s borders, Hodgson tracked wine that had been entered in at least two competitions in the United States and found that about 99% of the wines that earned gold medals in one event received no award some place else. Several gold medal-winning wines were entered in five competitions. None of them received five gold medals, or even four gold medals. In commenting on his research findings as a whole and their implications for wine judging events generally, Hodgson stated, "Chance appears to have a great deal to do with the awards that wines achieve or miss out on." He readily admitted that there are many individuals who are expert wine tasters with exceptional abilities to critically judge wines when a few samples are placed before them. Yet this is not what generally takes place at many wine-judging events. Hodgson explained, "When you sit 100 wines in front of a judge, the task is beyond anyone’s ability."

Hodgson’s findings received a great deal of media attention, with one of the more intriguing responses coming from Joe Roberts, a certified wine specialist who also was one of the judges at the California State Fair when Hodgson conducted his research. Roberts did not dispute or try to explain away the inconsistency in the judges’ ratings, but instead, argued that inconsistency is what defines a fine wine. On his wine blog, “1 Wine Dude,” Roberts stated the following:
“Fine wine should be changing in the bottle and in the glass. The wine I taste one minute should be different than the one I taste several minutes later, if the wine is any good. ... Put another way, do you know what could change a wine from a gold medal winner in one competition to a loser in another, even among the same judges? Anything. The barometric pressure, whether or not I had an argument with somebody, needed to take a dump, had a great song stuck in my head, ate a good breakfast, saw too much of the color red on billboard ads on the way into the judging hall that day, or got a pour into a glass that got polished with the wrong towel... It’s not that all competition judges suck at what they do, it’s that their task is handicapped into an artificial situation from the start. ... The system of quickly evaluating a wine isn’t natural, isn’t perfect, and isn’t simple, and so if our assumptions are wrong (e.g., humans have robot-like quality assessment ability, wine is static, etc.) then our conclusions based on the results are bound to be off, too.”

After all the research that I have conducted for this article, my humble opinion is that Roberts’ assessment of wine judging, in light of Hodgson’s findings, is “spot on.” I also believe that many wine judges would agree with his thinking, but what about those who actually organize such competitions? How did those who ran the California State Fair Wine Competition react to Hodgson’s findings? “In many instances, with mixed emotions,” said Hodgson. “Initially, some at the California State Fair Board tried to delay the release of my findings by advising, ‘Maybe next year when we have more data.’ Yet others were very supportive, stating ‘The more we know, the better we will understand.’ And then there were those who said, ‘If this information ever gets out, it will be the end of the State Fair wine judging!’” Anxiety was so great among some at the State Fair that just prior to Hodgson publishing his first set of findings he was asked not to identify the state in which the study took place. Initially, he agreed to this request, but an enterprising reporter at the Sacramento Bee quickly revealed where Hodgson’s data originated, so the cat was out of the bag. For the doomsayers who believed the sky would fall if the public learned about these findings, their fears were unwarranted. Today, the California State Fair Wine Competition is still very much alive and thriving.

**Are There Order Effects in Wine Judging?**

Another important question surrounding wine judging is whether wines receive different evaluations simply based on where they are placed in a wine-judging flight. This is a reasonable question because research dating back to the 1950s suggests that the order in which everyday wine consumers taste wines influences how much they like them. In 2009, for example, Brock University marketing professor Antonia Mantonakis published findings in the journal *Psychological Science* indicating that when people tasted a sequence of wines and were asked which wine they liked best—in reality, all wines were from the same bottle—they preferred the first wine when tasting just two or three glasses, but when tasting five or more glasses of wine they preferred the last wine. Mantonakis found that this shift in preference from the first wine tasted to the last wine tasted as the number of glasses in the sequence increased was most likely among the more experienced wine drinkers, a phenomenon in psychology known as the *recency effect*. The recency effect in decision-making is most likely to occur among people who have either been cautioned to weigh all the evidence before forming an opinion or who have learned to do so as part of their normal decision-making process. Generally, experts in any area are more likely than novices to reserve judgment in this manner.

![Professor Antonia Mantonakis](image)
Based on Mantonakis’ findings a natural question to ask is whether expert judges in wine competitions are susceptible to this same recency effect? In 2015 Food Science Professor Carole Honoré-Chedozeau and her colleagues at the University of Burgundy in Dijon published results of a three-year study that addressed this question. Participants in their study were approximately 100 wine professionals serving as judges in competitions that awarded medals to Beaujolais Nouveau wines in France. During the competition, each tasting judge was given two flights of between 10 and 12 wines; yet, in each tasting sequence the researchers inserted the same wine (a wine not registered in the competition) into the first and second-to-last positions. A comparison of the scores given by the judges to this “test wine” revealed that its average score was about three points higher when it was in the second-to-last position versus in the first position. This three-point swing was not only a statistically significant difference, it also would have resulted in more awards if these wines had truly been entered into the competition. Indeed, over the three annual wine competitions, a wine evaluated in the second-to-last position had 1.4 to 1.9 times greater likelihood of receiving a total score that qualified for an award than did the same wine when it was slotted in the first position for judging. Honoré-Chedozeau’s findings, along with those of Mantonakis, suggest that wines evaluated at the end of a flight of wines are more likely to receive higher scores—and thus, more awards—than are wines evaluated at the beginning. In explaining this apparent effect among experienced wine tasters, Mantonakis stated that these experts’ desire to make sure they identify the best wines causes them to fall prey to ultimately being more receptive to the last choices in the wine flights they judge. This is why the “sweet spot” for a wine in wine-judging competitions is at, or just near, the end of a flight.

Recommendations for Winemaking Clubs

Last year, inspired by Hodgson’s research, I conducted a pilot study during our Wisconsin Vintners Association’s (WVA) annual wine-judging event by clandestinely inserting two wines twice into two of our judging table flights. The most prominent inconsistency I found was that, of the two wines tested, judges gave a Zinfandel wine a 3rd place award the first time they tasted it and a 2nd place award a few minutes later. Judges at the two tables also gave different scores to the same wine about one third of the time in three of the five evaluation categories, namely, aroma and bouquet, taste and texture, and aftertaste. Faced with these judging inconsistencies, over the ensuing year the WVA Board of Directors implemented a wine-tasting education program for its judges and general membership. First we made a concerted effort to hold wine tastings at a number of our monthly meetings, each time teaching our members how to use our 20-point wine scoring system. We also offered a four-week intensive wine-evaluation course in which one-fourth of our members participated, with the final session featuring four of our association’s top wine judges offering their independent tasting evaluations for a series of wines, which attendees also evaluated. This comparison wine tasting, which used our club scoring system, was extremely illuminating because there were not only considerable differences observed between members’ evaluations of each wine, but also between those of our four expert judges. Indeed, there was not a single wine where the four judges’ individual independent scores fell into the same award category.
The best summary of this demonstration of judging consistency was offered by one of our four expert judges when he admitted at the end of the tasting, "Well, I guess we just proved the point of Hodgson’s wine-judging studies, didn’t we?"

Based on all the evidence presented in this article I think it is abundantly clear that no wine judge—regardless of their degree of experience—is infallible when it comes to evaluating wine, but my intention in presenting this critique is not to dismiss wine judging events as being without merit. Within amateur winemaking clubs, such events not only bolster the confidence of those who win awards, they hopefully also provide useful feedback for those whose wines fall short of the mark in the judges’ opinions. Each judge’s wine evaluation sheet contains useful feedback on important aspects of the wine’s components that can provide valuable guidance in future winemaking efforts. Further, if wine-judging events are organized properly, when judges perceive a significant fault in a wine, a separate sheet is provided the winemaker in which the wine fault is identified and advice is given on how to avoid similar problems in the future. Again, such feedback is where the true value lies in these events. Indeed, the term “event” is a more accurate descriptor for amateur wine judging than is the term “competition” because I’m guessing that the vast majority of individuals who enter wines for judging are not scouring out the other entrants as foes to be vanquished, but rather, as peers with whom they can collectively improve their skills.

Regarding amateur wine-judging events, it would be highly instructive for organizers to make it a regular practice to include into each judging table flight two identical “test wines” so that every judge receives feedback on their degree of judging consistency. Based on Hodgson’s research, we know that there is considerable inconsistency in judging wine, but Hodgson’s data further informs us that perhaps 10 percent of judges are extremely inconsistent and need replacing. Over time, these exceptionally poor judges can be discreetly identified and reassigned different roles, such as wine steward, wine score tabulator, or glass washer. Another bit of information that can be tracked at such events to possibly improve the validity of the judging is to keep tabs on whether judging tables differ in their assignment of 1st place, 2nd place, 3rd place, and “no award” judgments. If such differences are found, it is possible they reveal discrepancies in the leniency and harshness of the tables’ scoring standards. Another relevant and related question to track is whether there is a difference in what types of wine earn higher ratings at such events. Is it possible that fruit wines and grape wines are being judged by different quality standards? What about sweet versus dry wines? It is possible that those who make certain wine styles in a particular club are simply better winemakers than those who make other wine styles, but it is also possible that the source of this difference is the judges’ own personal preferences and quality standards. The goal here in all these recommendations is not to eliminate judging error; that is simply not possible. However, it is possible to improve wine judging by identifying a set of “fixable” errors.

One thought that some of you may have at this point is, “Why all this fuss? Does it really matter to have consistency in judging?” After Hodgson’s studies were first published, a friend of his who was also a world-renowned wine expert summarized this sentiment with the following statement: “Bob, this is just a wine competition; it’s not all that important. Get over it!” That declaration calls to mind one last study that I’d like to share. In a 2012 Journal of Wine Economics article, Duke University professor Robert Ashton compared the assessments of experienced wine judges to those of experts in the following six fields: medicine, clinical psychology, business, auditing, personnel management, and meteorology. Ashton’s findings were that in all fields,
including wine judging, some experts were better evaluators than other experts, but wine experts were overall substantially worse in their judgments than were experts in all other fields. To me, these findings are heartening, because I frankly consider judgment accuracy in medicine, psychology, and business much more valuable to my everyday living than wine judgment. It is often said that a wine review is a particular person’s impression of a wine at a particular moment in time. If you recall Joe Roberts’ earlier assessment of wine-judging events, then it is also safe to conclude that their evaluations are simply a set of impressions at a particular moment in time. If this blanket evaluation of wine judging seems reasonable, the next time you submit one of your wines to a competition, instead of getting wrapped up in whether it is assigned the award you were hoping for, you will have a better chance of improving your future winemaking efforts if you instead focus on the judges’ scoring sheet comments, weighing the wisdom of their feedback, with the full knowledge that perhaps some of those evaluations are not written in stone, but rather, shifting sand.

In keeping with the adage that “knowledge is power,” at our annual wine-judging events if entrants want to become better winemakers they should pay more attention to the judges’ comments on the wine-evaluation sheets rather than on whether or not they won an award.