


ARTICLE

Judging reliability at wine and water competitions

Elena C. Berg¹, Michael Mascha² and Kevin W. Capehart³ 

¹Department of Computer Science, Math, and Environmental Science, The American University of Paris, 5 Boulevard de la Tour-Maubourg, 75007 Paris, France, ²FineWaters, Fine Water Academy and ³Department of Economics, California State University, Fresno, 5245 N Backer Ave M/S PB20, Fresno, CA 93740

Corresponding author: Kevin W. Capehart, email: kcapehart@csufresno.edu

Abstract

Studies suggest the inter-rater reliability of judges at wine competitions is higher than what would be expected by random chance, but lower than what is observed when experts in other fields make judgments specific to their expertise. To further contextualize the (un-)reliability of wine judging while also extending the study of fine water, we examine the inter-rater reliability of judges at an annual international competition for bottled waters. We find that the inter-rater reliability of water judging is generally better than chance and, at best, about the same as the inter-rater reliability of wine judging at some wine competitions. These results suggest that perceptible differences between fine waters exist but are less pronounced than those between fine wines and, also, that aesthetic standards with respect to fine waters exist but are currently less established than those for fine wines.

Keywords: blind tasting; expert evaluation; inter-rater reliability; water; wine

JEL Classification: D83; L66; Q25

1. Introduction

Wine competitions—organized events in which a panel of experts blind taste wines, rate them on a numerical or other scale, and bestow awards on the highest rated entries—continue to attract attention, including from wine economists. A number of studies have examined the 1976 Judgment of Paris (e.g., Ashenfelter and Quandt, 1999; Ashton, 2011), the annual wine competitions at the California State Fair (e.g., Hodgson, 2008; Bodington, 2020), and other similar competitions (e.g., Bitter, 2017; Hodgson, 2009).

Those studies have evaluated, among other things, the inter-rater reliability of wine judging as measured by the correlation between different judges' ratings of the same wines. A positive correlation between judges' ratings suggests they identified similar differences among the wines and, moreover, shared similar aesthetic standards when translating perceived differences into ratings. In that case, their collective judgment about the relative quality of the wines might be meaningful.

Previous studies suggest the correlations among wine judges' ratings are generally positive and higher than what would be expected by random chance (Ashton, 2012;

Bodington, 2020). Wine judging, therefore, appears to have a higher inter-rater reliability than pure randomness. Previous studies also suggest that wine judging has a lower inter-rater reliability than what is observed when experts in other fields make judgments specific to their expertise (Ashton, 2012). However, some of the field-specific tasks to which wine judging has been compared—such as meteorologists forecasting hailstorms—seem only vaguely comparable.

This paper provides a new basis of comparison for the inter-rater reliability of wine judging. We analyze a field-specific task that is similar to wine experts blind tasting and rating wine at a wine competition. Specifically, we analyze water experts' blind tasting and rating of waters at an annual international bottled water competition called the "Fine Water Taste and Design Awards." The Fine Water competition is divided into two parts: the "Taste Awards," in which judges rate the blind taste of waters, and the "Design Awards," in which the same judges rate the visual design of bottles. Our primary focus will be on the Taste Awards, which is the most similar to a wine competition, but we will also briefly address the Design Awards as another potentially interesting basis for comparison.

Although all water might seem the same, even potable waters vary in terms of their total mineral content (which is measured by "Total Dissolved Solids" or TDS), mineral composition (such as their concentrations of calcium, magnesium, sodium, chloride, phosphorus, and silica), carbonation, pH, and other dimensions (as discussed by Capehart and Berg, 2018, and references therein). A number of studies, mostly conducted in the context of managing municipal water supplies to ensure public safety and satisfaction or understanding consumption patterns for bottled water, suggest that human taste buds are sensitive to variation in a water's TDS level and mineral composition, at least when mineral levels vary widely enough (Burlingame, Dietrich, and Whelton, 2007; Marcussen, Holm, and Hansen, 2013). Studies also recognize other ways in which humans can be sensitive to water's taste and odor (Dietrich and Burlingame, 2020) or health effects (Azoulay, Garzon, and Eisenberg, 2001).

The bottled waters that compete in the Fine Water competition are drawn from specific natural sources, as we discuss later. Each source's distinct geology can lead to distinct characteristics in its water. To the extent that the water experts judging the Taste Awards can identify similar differences among the waters they blind taste and, moreover, to the extent they share similar aesthetic standards, the inter-rater reliability of water judging should be better than random chance and could potentially be as high as that of wine judging.

We find that the inter-rater reliability of water judging at the Taste Awards is generally better than random chance, similar to wine judging at some competitions (including, in particular, the 1976 Judgment of Paris), and worse than wine judging at other competitions (including, in particular, a recent year of the California State Fair's commercial wine competition studied by Bodington, 2020). The fact that the Taste Awards have a lower inter-rater reliability than at least some wine competitions may be due to water experts having blind-tasting abilities that are less developed and aesthetic standards that are less established than wine experts. Such things could change over time if water expertise becomes as professionalized as wine expertise.

II. Background

A. Previous studies on the inter-rater reliability of wine judging

One measure of the inter-rater reliability between two judges' judgments is the Pearson product-moment correlation coefficient (Olkin et al., 2015). That measure, which is appropriate for cardinal data, has been used for several analyses of wine judging (e.g., Ashton, 2012; Bodington, 2020). A coefficient of (plus or minus) unity means there is a perfect (direct or inverse) linear association between the judges' judgments. Zero means no linear association. For intermediate values between zero and one in absolute terms, conventional rules of thumb are that a correlation larger than 0.10, 0.30, or 0.50 is "small," "medium," or "large," respectively (Cohen, 1992). Those rules of thumb are not unhelpful for interpreting intermediate values, but they are context-independent and somewhat arbitrary, so a richer contextualization would be preferable when possible.

Although a Pearson correlation coefficient can be used to compare the judgments of only *two* judges at a time, Ashton (2012), Bodington (2020), and others have compared the judgments of *more than two* judges by calculating a Pearson correlation coefficient for each possible pair of judges (i.e., pairing each judge with every other judge, except themselves) and then taking a simple average of all those correlation coefficients. Hereafter, we will refer to that as the "average pairwise correlation."

Ashton (2012) used average pairwise correlations to compare inter-rater reliability across several fields. He identified 46 studies in which experts from six other fields besides wine tasting made judgments specific to their field of expertise. Half of the studies (23 of the 46) were from the field of auditing, and half of the fields (three of the six, including auditing) were related to business (Ashton 2012, p. 79). For most of the studies of experts in a given field, Ashton was able to obtain an average pairwise correlation for their judgments.

Ashton (2012) did the same for the field of wine tasting, identifying nine studies in total. Two of those studies were Ashton's own earlier analyses of the white wine flight and, separately, the red wine flight at the 1976 Judgment of Paris (Ashton, 2011). Two more studies were analyses of those same flights at that same competition, but using an intraclass (rather than Pearson) correlation coefficient. Four of the nine studies were analyses of lab experiments set up similarly to wine competitions. The remaining study was Hodgson's (2009) study of awards won at different wine competitions (rather than judges' ratings from one wine competition). Thus, only two of the nine wine-tasting studies identified by Ashton (2012) were actually studies of the pairwise Pearson correlation coefficients for judges' ratings at a wine competition. We will nevertheless presume in the discussion that follows that Ashton's (2012) results are representative of the Pearson correlation coefficient that would be expected for wine experts' ratings, as well as the other experts' field-specific judgments.¹

¹Ashton's (2012) findings could obviously change if other studies found different levels of inter-rater reliability than the studies he identified. While discussing Ashton's (2012) findings, Oczkowski (2017, p. 58) identifies four additional studies of inter-rater reliability for wine tasting that report an average pairwise Pearson correlation coefficient. None of those studies were about judges' ratings at wine competitions; instead, they were about the published ratings of famous wine critics such as Robert Parker. Despite that, it could still be the case that there are other published studies (or, due to publication biases, unpublished

Table 1. Ashton's (2012) findings for inter-rater reliability in various fields

Field	Number of studies	Example of field-specific task	Inter-rater reliability
Meteorology	4	Forecasting hailstorms	.75
Personnel management	6	Evaluating employees for promotion	.65
Auditing	23	Making financial reporting judgments	.61
Medicine	3	Evaluating the severity of patients' ulcers	.56
Business	8	Analyzing businesses' tax liabilities	.49
Clinical psychology	2	Evaluating patients' sociability	.37
Blind wine tasting	9	Rating Californian and French wines	.34

Notes: This table is essentially the same as the table from Ashton (2012, p. 79). This table shows a measure of inter-rater reliability of experts in a given field when they make judgments specific to their field of expertise. Inter-rater reliability is measured as a simple average of the average pairwise Pearson correlation coefficients (or, in at least two cases mentioned in the main text of this paper, an intraclass correlation coefficient) reported by the studies identified by Ashton (2012). The examples of field-specific tasks are paraphrased from his paper (p. 78).

Ashton's (2012) findings, summarized in Table 1, suggest that the average pairwise correlation for wine experts' ratings of the blind taste of wine is greater than zero (specifically, 0.34) and, as such, at least some of their pairwise correlations must be greater than zero, too. An obvious question is whether the average correlation or any of the pairwise correlations are statistically significantly greater than zero. Ashton (2012) did not consider statistical significance, so we reanalyzed the studies he identified by following Bodington (2020). Bodington (2020) showed that, for the vast majority of judges at a recent year of the California State Fair commercial wine competition, the pairwise Pearson correlation coefficients for their ratings were greater than what would be expected by random chance alone. We confirmed that the correlations among the wine experts' ratings at the Judgment of Paris and the lab experiments identified by Ashton (2012) were also generally greater than what would be expected by chance (see the supplementary appendix to this paper for a full discussion).

Ashton's (2012) findings also suggest that, although the average pairwise correlation between wine experts' ratings is greater than zero (again, 0.34 by his calculations), it is relatively low. It is lower than the reliability with which meteorologists forecast hailstorms, or personnel managers evaluate employees for promotion, or auditors judge financial reports, to name a few examples.

Yet it is not clear that the field-specific tasks he considered are entirely comparable to wine judging. Ashton himself (Ashton, 2012, p. 82) recognized that wine judging is not entirely comparable to the other field-specific tasks and, moreover, that how it

studies) of wine judging or other field-specific tasks with different average pairwise correlations than those reported by Ashton (2012).

differs may lend itself to lower levels of inter-rater reliability. He points out that wine judges must rely on their own sight, smell, and taste senses rather than more objective instruments. Translating any given sensory experience into a rating would also involve some subjectivity. Bitter (2017, pp. 396–397) makes similar points about the difficulty and subjectivity of wine judging. It is therefore of interest to compare wine judging to something more similar.

B. Fine Water Taste and Design Awards

As discussed by Capehart and Berg (2018) and Capehart (2015) in this journal, there is a world of fine bottled waters akin to the world of fine wine. In addition to water sommeliers, water menus, and water guidebooks, there are also water competitions. One such competition is the annual “Fine Water Taste and Design Awards,” an international bottled mineral water competition run by the Fine Water Society, which in turn is run by the water expert Michael Mascha. Mascha has published a guidebook to bottled waters (Mascha, 2022), operates a website at <*finewaters.com*> with similar information, co-operates a program called the Fine Water Academy that trains and certifies water sommeliers (Fine Water Academy, 2022), and overall, actively promotes fine bottled waters (Hooks, 2013). Mascha has run the Taste and Design Awards five times in total as of this writing (specifically, from 2016 to 2019 and 2021, with a planned 2020 competition canceled due to the pandemic). Bottled water companies from all over the world have competed, and the competition has been held in four different cities around the world (Guangzhou, China in 2016 and 2017; Machachi, Ecuador in 2018; Stockholm, Sweden in 2019; and Bled, Slovenia in 2021).²

The Fine Water competition is divided into two separate parts: the Taste Awards and the Design Awards. Both parts are judged by the same panel of five judges, who are all considered to be water experts. Promotional materials for each competition include short biographies for each judge highlighting their training, experience, or other water-related credentials (see, e.g., Fine Water Society, 2021).

Another notable water competition is the Berkeley Springs International Water Tasting, which has been held annually for about 30 years in Berkeley Springs, West Virginia. We will not study that competition here, mostly for the practical reason that we have been unable to obtain judge-level data from the organizers of that competition. A more principled reason to ignore that competition is that at least some of its judges are novices who receive minimal training before tasting (Fulcher, 2017). To compare our results to wine competitions, which usually involve wine experts, we opted to look at a water competition where all of the judges are water experts. There is evidence, discussed later, that novices who blind taste and rate bottled waters have a much lower inter-rater reliability than the judges at the Taste Awards.

To compete in the Fine Water Taste and Design Awards, a company has to supply its own bottles and pay a monetary fee if they are not already a member of the Fine Water Society. To be accepted, water must come from a natural source and be

²A precursor of sorts to the Fine Waters Taste and Design Awards was held in 2015 as part of a water industry exhibition in Guangzhou, China; but again, the first competition under the auspices of the Fine Water Society was in 2016 (Fine Water Society, 2016).

unprocessed as understood by the Fine Water Society. Acceptable natural sources include springs, wells, rain water, glacial runoff, and icebergs, as well as sources that are considered more exotic, such as the deep sea. The water also must be as unprocessed as possible under the regulations of the source country, where micron-filtering, ozonation, or ultraviolet treatment are often required. The desire is to have the water as unaltered as possible from its original form, with two exceptions: carbonated waters and “curated” waters. To create sparkling water, it is permitted to add carbon dioxide, although there is a separate category for “naturally” sparkling waters because carbonation can occur naturally under certain rare geological conditions. “Curated” waters consist mainly of *cuvées*, where water from multiple natural sources is mixed, and single-sourced waters, where certain minerals are added in order to create a particular desired taste profile. “Processed” waters, including municipal tap water or any reprocessed versions of that water, are not allowed to compete. This is in contrast to the Berkeley Springs competition, which allows municipal waters and bottled waters to compete, albeit in separate categories. The Berkeley Springs competition also has a category for what they call “purified” water (Fulcher, 2017), which is an example of what the Fine Water competition would call processed water.

C. More on the Taste Awards

For the Taste Awards part of the Fine Water competition, the five judges are seated side by side at a table. They face an in-person audience (which was restricted in the Covid-impacted years) and a camera for live streaming the event to a virtual audience. On the table in front of each judge, there is a clear tasting glass, a sheet of paper with an empty table corresponding to each category of water being tasted, a writing instrument for recording scores, and a small flip scoreboard that faces out towards the audience. The scoreboard is used to display integer values from 90 to 100, which is the scale on which the judges score the taste of the water.

When the blind tasting starts, an attendant brings out a bottled water in its original bottle but covered to the top of the neck in a dark bag that masks any identifying information about it. All waters are opened immediately before they are poured and are tasted at room temperature, with care taken to ensure that all waters are at the same temperature. The attendant approaches the first judge on stage right from the judge’s right, pours some water into their glass, and then repeats this process for the other judges. Once a judge receives a pour, they taste the water, record their score on the sheet of paper in front of them, flip the scorecard to show the audience that score, and discard any remaining water into a dump bucket beside them. Once the attendant has poured water for each judge, the attendant moves around to the front of the table, removes the bag, and places the bottle on a low table with its label facing the audience. The table is far enough below the judging table that the bottles placed there are not visible to the judges. The bottles on that low table are visible to the audience, so they can see the brands of water being judged.

It should be recognized that because the judges are not completely isolated from each other, they could influence each other’s ratings. Yet that could be true for wine competitions, too. At the Judgment of Paris, participants sat beside each other and

even sometimes talked openly about the wines (Taber, 2006, pp. 200, 203). At the California State Fair commercial wine competition, judges sit beside each other (see, e.g., Cal Expo, 2019). That competition is atypical because judges can confer with each other about their individual ratings before awards are assigned (Hodgson, 2008, p. 106). At the Taste Awards, no such conferral occurs. For the purposes of the analysis, we treat the water judges' ratings as independent.

The previously-described process repeats itself one bottle at a time, all in one sitting. The pace of the tastings is quick, with the entire event typically spanning a little over an hour, depending on the number of entries. There have been as few as 42 waters tasted in 2016, as many as 87 in 2019, and about 66 waters tasted on average per year over the five years of the competition. Each water is served only once in a given year; there are no replicated samples.

The Berkeley Springs bottled water competition also asks its judges to evaluate a large number of waters; they evaluate four flights of 20 waters at a time over the course of one day (as noted by Capehart and Berg, 2018, p. 25). Judges at some large wine competitions evaluate a similarly large number of wines in a day (50 or so) and do so multiple days in a row (as noted by Bodington, 2020, p. 364, about the California State Fair's commercial wine competition).

Waters are grouped into categories that are tasted in a certain order. Within each category, the waters are tasted in an order randomized by a master of ceremonies who is not one of the judges. The first category is still waters, with TDS levels in a relatively low range. Still waters with TDS levels in increasingly higher ranges are tasted in subsequent categories. After that, sparkling waters are tasted, also moving from low to high TDS levels. See Mascha (2022) for his preferred demarcation of TDS levels. The naturally sparkling waters, curated waters, and waters from exotic sources are separate categories.

After all the waters have been blind tasted, the score sheets are removed from the table for entry into a spreadsheet. At the end of the day, the judges' scores are averaged and gold, silver, and bronze medals are awarded to the highest-rated waters in each category.³

D. More on the Design Awards

Immediately before judging the Taste Awards, the same five judges judge the Design Awards. For that part of the competition, the judges look at bottled water containers and rate their visual design on a 90- to 100-point scale. The bottles are arranged on a long display table. They are organized into categories based on the material out of which they are made. Categories have included glass containers, polyethylene terephthalate (PET) plastic containers, aluminum cans, and Tetrapak cartons. No aspect of the visual design or label of a bottle is masked from the judges in any way. The judges may pick up or touch the containers if they would like to do so, but they

³When the competition averages the judges' scores to assign awards, they ignore the score a judge gives to a brand if the judge has a financial conflict of interest with that brand. So, for example, if a judge is the CEO of a company whose brand is competing in the competition or a consultant to the brand, then that judge's score for that brand is ignored when averaging the judges' scores. In our analysis, we use all the judges' scores from the Taste Awards and do not attempt to deal with any concerns about strategic rating.

do not open the bottles or taste any of the waters at this time. Each judge records their scores for the visual design of the bottles on a sheet of paper without conferring with the other judges. As with the Taste Awards, the judges' scores are entered into a computer spreadsheet; the scores are averaged; and gold, silver, and bronze medals are awarded to the highest-rated bottles in each category. Neither the Design nor the Taste Awards are announced until the end of the day. Most, but not all, of the bottled waters that compete in the Design Awards also compete in the Taste Awards.

III. Data and methods

The main contribution of this paper is to compare the inter-rater reliability of the water experts' ratings at the Taste Awards, as well as the Design Awards, to the inter-rater reliability of wine experts' ratings at wine competitions. The authors of this paper were able to obtain the judge-level ratings for all five years of the Fine Water competition. In principle, the data for the Taste Awards could be reconstructed by watching the playback recordings that are publicly available through the previously-mentioned *<finewaters.com>* website, although it is easier to have that data provided by the organizers of the competition rather than independently reconstructing it. There are no playback recordings of the Design Awards, so only the organizers of the competition have a record of the judge-level ratings. The organizers of the competition provided the data we used without any formal agreement about what results could (or could not) be published based on the data.

In total, for the five years' worth of data we obtained, we have 1,645 blind-tasting-based ratings of 190 unique bottled waters from the Taste Awards. We also have 1,230 ratings of the visual design of 155 unique bottled waters from the Design Awards.⁴ Those ratings came from 16 unique judges over the years. Eleven judged during a single year only (including three who were new in 2021), three judged during two years, one judged in three years, and one (Mascha) has judged all five years.

To examine inter-rater reliability between those judges' ratings, we will follow Ashton (2012), Bodington (2020), and similar studies by using the Pearson correlation coefficient. It should be recognized that, if judges' ratings are or should be treated as ordinal (rather than cardinal), then another measure of inter-rater reliability such as the Spearman rank-order correlation coefficient should be used. The Spearman correlation coefficient is equivalent to a Pearson correlation coefficient over ranks; it measures whether there is any monotonic (rather than simply a linear) relationship between two judges' judgments. Ashenfelter and Quandt (1999), Gergaud, Ginsburgh, and Moreno-Ternero (2021), and others have argued in favor of using ordinal rankings when analyzing wines and wine judging. That said, the results

⁴In three instances for the 2021 Design Awards competition, we have a record of only the average score that each judge assigned to two different bottles of the same brand of bottled water (rather than the scores each judge gave to each bottle from that brand). Here and throughout, we treat those averages as if they were the scores assigned to a unique bottled water. For five bottles competing in the 2021 Design Awards competition, we do not have a record of a judges' score. No score was recorded when a judge had a financial conflict of interest with a brand. We ignore those five bottles when analyzing the Design Awards.

presented later are similar if the Spearman (rather than Pearson) correlation coefficient is used; see the supplementary appendix to this paper for those similar results.

A Pearson (or Spearman) correlation coefficient could be calculated whenever any two judges rate at least two items.⁵ Yet if a correlation coefficient is calculated over only a small number of items, then that can limit the possible correlations (in the extreme case of only two items, the only possible correlations are positive or negative unity) and spurious correlations are more likely. For the Taste and Design Awards, the number of bottled waters in any given category can be quite small (as little as one bottle for some categories in some years), and it is obviously smaller than the number of bottled waters across all categories. We will therefore calculate a Pearson correlation coefficient across all the categories of bottled water rated when examining the Taste Awards and, separately, the Design Awards.

In addition to using the Pearson correlation coefficient, we will follow Bodington's (2020) analysis of the 2019 California State Fair's commercial wine competition in four other ways. First, we will examine not just the average pairwise correlation between judges' ratings (as in Ashton, 2012), but also the entire distribution of pairwise correlations. We will do this graphically by drawing "violin plots" of the pairwise correlations for judges' ratings at a given competition. The violin plots show the minimum, interquartile range, median, and maximum of the distribution of pairwise correlations for judges' ratings at a given competition, as well as kernel density estimates of that distribution.⁶ Bodington (2020, p. 366) draws a density estimate (specifically, a histogram) for the distribution of pairwise correlations for the 2019 California State Fair wine competition.

Following Bodington (2020), we will also consider the overlap (or lack thereof) between different distributions of pairwise correlation coefficients. In addition to visually inspecting the violin plots we will draw, we can quantify this overlap by using the "probability of superiority" (PS) statistic. For any two distributions of pairwise correlation coefficients, the PS of one over the other is the probability that a randomly selected correlation from the former distribution is greater than a randomly selected correlation from the latter distribution. If all of the correlations in the former were greater than all of the ones in the latter, then the PS would be 100%. The PS would be 50% if the distributions were identical. We will use the Mann-Whitney U test in order to test the null hypothesis that the PS for two distributions is 50%.⁷

Yet another way in which we will follow Bodington's (2020) analysis is to compare the distribution of pairwise correlation coefficients that was *actually* observed for a given competition to the distribution of pairwise correlation coefficients that *would have been* observed if a competition had been set up exactly like the given competition (in terms of the number of judges, the number of items assessed, and the possible ratings), except that each judge was randomly rating each item by drawing from a

⁵The variance of each judge's ratings also has to be non-zero, or else the Pearson (or Spearman) correlation coefficient is undefined.

⁶For the kernel density estimates, we use Gaussian kernels and the Silverman bandwidth rule, although the figure shown would look similar using other kernels or bandwidths. We truncate the distributions at the minimum and maximum pairwise correlations in order to show the observed range of pairwise correlations. For an accessible explanation of kernel density estimation, see DiNardo and Tobias (2001, pp. 13–20).

⁷For more on the PS statistic and its relation to the Mann-Whitney U test statistic, see Ruscio (2008).

uniform distribution over the possible ratings. If the former distribution has little overlap with the latter distribution, then we conclude that the observed pairwise correlations are generally greater than what would have been expected by chance. Bodington (2020) does the same to assess whether the inter-rater reliability of judging at the 2019 California State Fair wine competition was any better than chance.

We will also follow Bodington's (2020) analysis by using the same data he obtained for the 2019 California State Fair commercial wine competition as one of our representations of the inter-rater reliability for wine judges' ratings.⁸ For that wine competition, there were 18 panels of three judges apiece. Each judge had to complete training and pass a test of their discriminatory tasting abilities (Bodington, 2020, p. 364). Each panel of judges tasted more than 100 wines over two days. The wines were grouped into roughly five flights per day, with roughly 10 wines per flight. Each flight had wines of a given type, such as Chardonnays or generic reds. A small number of hard ciders (seven in total) were also judged in some flights, but we ignored those non-wines in our analysis. For a given wine, each judge rated the wine as either "Gold+," "Gold," "Gold-," "Silver," "Bronze," or "No Award." We convert those ratings to a six-point scale, following Bodington (2020). We calculate pairwise Pearson correlation coefficients over all the wines that judges on a given panel judged over two days (rather than over each flight or other narrower grouping) so that those correlations are more comparable to the ones we calculate for the Taste and Design Awards.

The 2019 California State Fair commercial wine competition may not be representative of other wine competitions, so we will also analyze the famous and much-analyzed 1976 Judgment of Paris. We examine these two competitions not only because we had access to judge-level ratings, but also because they exhibit different levels of inter-rater reliability, as detailed later. They do not necessarily cover the entire range of inter-rater reliability observed across wine competitions. Indeed, another competition we look at in the supplementary appendix, the 2012 Judgment of Princeton, exhibits a lower level of inter-rater reliability than any of the ones examined here. Nevertheless, we hope that the two we have selected are somewhat representative of wine competitions with relatively low and high levels of inter-rater reliability.

For analyzing the Judgment of Paris, we use all the data reported by Hulkower (2009).⁹ Similar to our analysis of the Taste Awards, Design Awards, and California State Fair, we will calculate pairwise Pearson correlation coefficients over all 20 wines (rather than over the white and red wine flights separately) for each possible pair of all 11 judges.

IV. Results and discussion

Figure 1 illustrates the main results of our study. It shows violin plots of the pairwise Pearson correlation coefficients for the judges' ratings at the four different competitions we consider: the Taste Awards competition for the blind tasting of bottled

⁸He obtained that data through a Freedom of Information Act request. The ratings provided by the California State Fair were for the judges' *initial* rating of a wine (before they conferred with each other).

⁹As Hulkower (2009) notes, there is some debate about whether the scores of the two non-French judges should be considered and whether the judge-level white wine scores are actually the original ones; but again, we use all the data he reports.

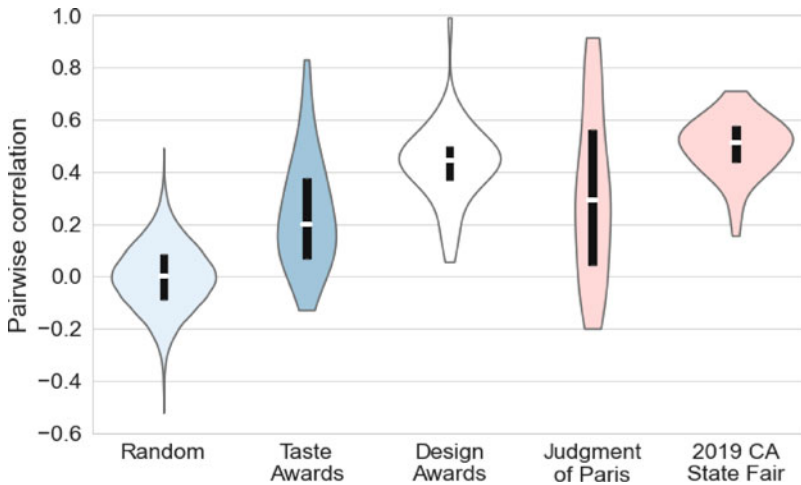


Figure 1. Violin plots of pairwise Pearson correlation coefficients.

waters; the Design Awards competition for the visual design of bottled water containers; the 1976 Judgment of Paris competition for Californian and French wines; and the 2019 California State Fair commercial wine competition. We will discuss the final violin plot, labeled “Random,” next.

A. Taste Awards are more reliable than random chance

As shown in the violin plot labeled “Taste Awards” in Figure 1, the distribution of pairwise correlations for the judges’ blind-tasting-based water ratings of bottled water over the five years of the Taste Awards has a median of 0.20 (as indicated by the white line in the middle of the violin plot), an interquartile range of 0.07 to 0.38 (as indicated by the edges of the black bar in the middle of the violin plot), and a range from as low as -0.13 to as high as 0.83 (as indicated by the range of the kernel density estimates). The average pairwise correlation is 0.24. About 90% of the pairwise correlations are at least slightly (though not necessarily statistically or substantively) greater than zero.

To assess whether the pairwise correlations from the Taste Awards are any better than what would be expected by chance, we simulated the pairwise correlations that would have been observed if a competition had been set up exactly like the Taste Awards (with five judges per year, the same number of waters tasted as those that were actually tasted in each year, and each water rated on a 10-point scale), except each judge’s rating for each water was an independent draw from a uniform distribution over the integers from 90 to 100. A violin plot of that simulated distribution is shown in Figure 1 and labeled as “Random.” The median and mean of that simulated distribution are essentially zero, as they should be. Of interest are the non-zero pairwise correlations that happen by chance.

Although some of the simulated pairwise correlations are as large as those observed for the actual Taste Awards, the correlations for the Taste Awards are

generally greater than the randomly generated ones. The PS for the distribution of the correlations from the actual water tasting over the distribution of randomly generated correlations is about 84% (p -value < 0.01 based on the previously-mentioned Mann-Whitney U test with a null hypothesis that the PS is 50%). We therefore conclude that, like at least some wine competitions, the inter-rater reliability of the judges' ratings for the Taste Awards is generally better than what would be expected by chance.

That finding contrasts with the results of Capehart and Berg's (2018) study of novices (rather than expert) water tasters. They had over 100 water novices (specifically, a sample of undergraduate students) rate the blind taste of four bottled waters and a local tap water. A reanalysis of their data shows that the inter-rater reliability among those novices is indistinguishable from what would be expected by chance (as we detail in the supplementary appendix to this paper). The judges at the Taste Awards may therefore have some expertise that novices lack.

B. Taste Awards are as reliable than some but not other wine competitions

The violin plots labeled "Judgment of Paris" and "2019 CA State Fair" in [Figure 1](#) show the pairwise correlations for the Judgment of Paris and the 2019 California State Fair commercial wine competition, respectively. For the correlations from the Judgment of Paris, the median is 0.29, the interquartile range is 0.04 to 0.56, and the range is -0.19 to 0.92. The average is 0.31, which is similar to the average pairwise correlation of 0.34 that Ashton (2012) reports for the nine wine-tasting studies he identified (where, again, the only two studies that were actually studies of the pairwise Pearson correlation coefficients for judges' ratings from a wine competition were Ashton's own earlier studies of the white and red wine flights from the Judgment of Paris).

The PS for the pairwise correlations from the Judgment of Paris over the correlations from the Taste Awards is only about 56% and not statistically significantly different from 50% at conventional levels (p -value = 0.33). Wine competitions and the Taste Awards are therefore fairly similar in terms of their inter-rater reliability if the Judgment of Paris is any indication.

However, as can be seen in [Figure 1](#), the 2019 California State Fair wine competition has a higher inter-rater reliability than either the Judgment of Paris or the Taste Awards. For its correlations, the median is 0.51, the interquartile range is 0.44 to 0.58, and the range is 0.15 to 0.71. The average is 0.50. The PS for its correlations is 85% relative to the Taste Awards' (p -value < 0.01) and 71% relative to the Judgment of Paris' (p -value < 0.01).

Without knowing more about the wines, waters, judges, or any other relevant aspects of those competitions, we can only speculate on why inter-rater reliability is higher for some wine competitions compared to other wine or water competitions. Inter-rater reliability would seem to vary as a function of: (1) objective differences among the beverages being judged (such as differences in wines' compounds, waters' minerals, and the like), where larger differences should be easier for any given judge to identify; (2) the ability of the judges to identify given differences, where part of the expertise of an expert should be the ability to more easily identify any given

difference; and (3) the extent to which the judges share similar aesthetic standards for translating identified differences into ratings.¹⁰

In terms of the first of those three potential sources of variation in inter-rater reliability: Some competitions may indeed have wines or waters that are easier to distinguish than those at other competitions. For the Judgment of Paris, the Californian and French wines that competed were selected because they were some of the best that their respective regions had to offer (even if the French wines were still expected to be better than the Californian ones; Taber, 2006). Perhaps the wines at that competition were too equally good to easily distinguish. The same might be true for the 2012 Judgment of Princeton, where some of the best wines from New Jersey and France competed (Taber, 2012), and inter-rater reliability was even lower than in the Judgment of Paris (as mentioned earlier). In the California State Fair commercial wine competition, in contrast, any wine made from Californian-grown grapes can be entered. Perhaps the wines in that competition are more different from each other and thus easier to distinguish.

Similarly, the waters at the Taste Awards might be at least as difficult to distinguish as the wines at some wine competitions. Even water experts admit that waters have subtler taste differences than wines (Mascha, 2022). Still waters (which obviously do not vary in their level of carbonation) and waters with relatively low TDS levels (which provide less to taste in terms of mineral concentration) would seem to be especially difficult to distinguish from each other. There is some evidence, reported in the supplementary appendix, that those waters are indeed more difficult to distinguish from each other compared to waters with carbonation or higher TDS levels. The fact that the waters at the Taste Awards are all fine ones may also be part of the explanation for that competition's relatively low inter-rater reliability. Perhaps if the Taste Awards allowed "unfine" waters to compete against fine ones, then that could lead to an inter-rater reliability as high as any wine competition. To take an extreme example, just like almost anyone could distinguish an extremely vinegary wine from most other wines, extremely chlorinated water could be easily distinguished from most tap or bottled waters.

In terms of the second potential source of variation: Even if we assume that judges at a wine or water competition have expertise that separates them from non-experts, different competitions could still vary with respect to their judges' ability to identify objective differences. The California State Fair and some other current wine competitions require judges to pass tests of their discriminatory tasting abilities (Bodington, 2020, pp. 363–364). Although that by itself does not necessarily mean the judges at the Judgments of Paris and Princeton or the Taste Awards had discriminatory abilities that fell short of the judges at the California State Fair, differences in those abilities are another possible explanation for our findings.

To the extent that the judges at the Taste Awards do fall short, it could perhaps be because water experts do not seem to be preoccupied (at least not yet) with blind

¹⁰In truth, it may not be possible to disentangle seemingly objective differences among items, on the one hand, and subjective judgments about the items, on the other hand, when the only instruments being used to judge differences are our naked human senses. However, for the purposes of the discussion that follows, it is useful to assume they can be disentangled.

tasting a water and trying to floridly describe its characteristics or guess its provenance. As evidence of that, the final examination for the Fine Water Academy's certified water sommelier program does not involve deductive blind-tasting tests like those required for some wine sommelier programs (Fine Water Academy, 2022). As further evidence, tastings led by water experts are typically unmasked and mostly involve discussing objective aspects of a water such as its source and minerals (see, e.g., Riese, 2022). Water menus curated by water sommeliers also seem to focus on those more objective aspects and say little directly about the sensory experience (Biro, 2019).

In terms of the third potential source of variation: A high inter-rater reliability requires that there are objective differences between the beverages being judged, that judges identify those differences, and that judges translate identified differences into ratings in a way that is consistent not just within but also between judges. Note that to ensure a strong positive correlation between judges' ratings, it is not enough for each judge to have their own internally consistent way of translating differences into ratings; in that case, there could be a strong *negative* correlation between some judges' ratings (if one judge rates highly what another judge rates lowly) or no correlation between their ratings (e.g., if the most salient aspects of a beverage for some judges are different from the most salient aspects for other judges and if those aspects are uncorrelated across beverages). For their scores to be highly correlated with each other, judges must share similar aesthetic tastes. To the extent that the judges at the Taste Awards have more disparate tastes than judges at some wine competitions, it could perhaps be because aesthetic standards are not as developed and established in the world of fine water as they are in the world of fine wine.

It is often said that learning about wine involves developing one's palate or refining one's tastes. There is also some evidence that wine experts have not only different abilities but also different preferences than novices (Ashton, 2017). At least part of becoming a wine expert would therefore seem to involve being inculcated, enculturated, or otherwise led or drawn towards certain aesthetic standards. There is by no means complete consistency among wine experts with respect to how they judge certain qualities of wine (e.g., where opinions differ on *Brett*) or the overall quality of wine (such as whether a wine is outstanding or poor, or worthy of an award or a particular number of points on some scale), which can perhaps explain at least part of the inter-rater un-reliability at wine competitions. Yet there is not complete inconsistency, either. Much of the world of wine—from training and certification programs offered by organizations such as the Wine & Spirit Education Trust or Court of Master Sommeliers, to the reviews of Robert Parker and other wine critics who write for outlets such as *Wine Advocate*, *Wine Enthusiast*, and *Wine Spectator*, to wine competitions—can be seen as reflecting and affecting aesthetic tastes.

Within the world of fine water, in contrast, there seems to be less of a culture (at least for now) of group-licensed or self-styled experts trying to describe or prescribe which fine waters or qualities thereof are more or less favorable. The Fine Water Society can be seen as trying to create differentiation between natural unprocessed waters and everything else, but with the notable exception of its bottled water competition, it does not seem preoccupied with creating a hierarchy among fine waters. Martin Riese—who has judged three of the five years of the Fine Waters competition, co-operates the Fine Water Academy, and is perhaps the most widely known water sommelier today

thanks in part to his appearances in traditional media and his activity on social media—often says there is no “best” water (while also asserting that natural unprocessed bottled waters are better than processed bottled waters; see, e.g., Tishgart, 2017). There is no equivalent of a Robert Parker of water who gives slightly different ratings to mostly similar waters on some 100-point scale. The bottled water guidebook and associated website mentioned previously are perhaps the closest things to anything like a *Water Advocate*, *Water Enthusiast*, or *Water Spectator*. The Fine Water Academy is as close as it gets to a Water & Education Trust or Court of Master Water Sommeliers. All of that could change, of course, which is a point we return to in the conclusion of this paper.

C. Design Awards are more reliable than Taste Awards and similar to wine competitions

Our main interest is in comparing the Taste Awards to wine competitions, but Figure 1 also shows a violin plot of the pairwise correlation coefficients for the Design Awards. The smallest and largest pairwise correlations are 0.06 and 0.99, respectively. The interquartile range is 0.37 to 0.50, the median is 0.44, and the average is 0.43.

The PS for the distribution of pairwise correlations from the Design Awards over the correlations from the Taste Awards is about 78% (p-value < 0.01). We therefore conclude that the inter-rater reliability with which the judges rate the visual design of bottles is higher than that with which the judges at the Taste Awards rate the blind taste of water. That finding is not too surprising, given that seeing the visual design of a bottled water would not seem to require the sort of keen sensory skills required for blind tasting. Also, bottled waters vary widely in their visual design because there are no fixed standards for their shape, size, or labeling. Nonetheless, it is intriguing that the judges apparently identified similar differences in the visual design of bottled waters and shared similar aesthetic standards for what a good-looking bottle looks like, more so than they were able to identify similar differences in the blind taste of waters and/or share similar aesthetic standards for what a good-tasting water tastes like.

The pairwise correlations for the Design Awards are somewhat larger than those for the Judgment of Paris (with a PS for the former over the latter of 65%; p-value < 0.01), but somewhat smaller than those for the 2019 California State Fair (with a PS for the latter over the former of 67%; p-value < 0.01). Thus, overall, the inter-rater reliability for the Design Awards seems fairly similar to that for wine competitions. The explanation for that finding would again seem to turn on the extent to which it is easy or difficult to identify differences among the items they are rating and, also, the extent to which the judges share or fail to share aesthetic standards. Perhaps it is more difficult for wine experts to identify differences in the blind tasting of wines than it is for the judges at the Design Awards to identify differences in the visual design of bottles, yet perhaps the wine experts are more consistent among themselves in terms of translating the differences they do identify into ratings.

IV. Conclusion

Our main conclusion is that the inter-rater reliability of water experts blind tasting and rating fine waters for the Taste Awards is generally better than what would be expected by random chance alone, but it is at best about the same as the inter-rater

reliability of wine experts blind tasting and rating fine wines at some wine competitions. The fact that inter-rater reliability at the Taste Awards is better than chance suggests that, to some extent, the fine waters that have competed have been different, the judges have been able to identify those differences, and they translated the differences into ratings in a way that was consistent within and between judges. Similar statements could be made about any wine competition with an inter-rater reliability of better than chance.

Yet the fact that the Taste Awards' inter-rater reliability was roughly the same as some wine competitions and worse than other wine competitions suggests one or more of the following may be true. First, differences among fine waters at the Taste Awards might not have been as pronounced as differences among the wines at some wine competitions. Second, judges at the Taste Awards might not have been as skilled as judges at some wine competitions in terms of identifying any given differences in the beverages they were blind tasting. And finally, compared to judges at some wine competitions, judges at the Taste Awards might not have been as consistent in translating identified differences into ratings, perhaps because aesthetic standards are less established for fine water than for wine.

Although this study was only able to consider *inter*-rater reliability, if it were possible to examine *intra*-rater reliability at the Taste Awards or another similar water competition, then that would yield insight into whether each judge can repeatedly identify the same qualities in a water and translate those qualities into a rating in a way that is internally consistent. If the *intra*-rater reliability of water judges were shown to be as *high* as that of wine judges, that would suggest that disparate tastes among water judges (rather than a lack of objective differences between waters or an inability of water judges to identify objective differences) is the main reason why water competitions do not compare more favorably to wine competitions in terms of their inter-rater reliability. If the *intra*-rater reliability of water judges is *low*, then there must be a lack of potentially perceivable differences among waters or a lack of perceptible ability among judges.

If water expertise continues to professionalize in a way that mirrors wine expertise (with a Court of Master Water Sommeliers, a Robert Parker of water, and the like), then water experts might continue to refine their blind-tasting skills and develop a more widely shared aesthetic system. Fine water competitions may also see greater variation among waters as that industry continues to expand. Allowing fine and unfine waters to compete in the same competition would also presumably allow for greater variation, as suggested earlier. Of course, if an unfine water ever were to beat out a fine water, then the consequences for the water industry could be as dramatic as those of the Judgment of Paris for the wine industry.

Another, different direction in which the world of fine water might evolve is to focus on other dimensions in addition to the blind taste of bottled water or the visual design of its container. Given widespread concern that the bottled water industry contributes to plastic pollution, unsustainable extraction of fresh water from its natural sources, unwarranted distrust of municipal tap water, and other environmental and social ills (see, e.g., Gleick, 2010), water experts could perhaps become experts on exactly how bottled water impacts the environment and society. A similar trend might already be emerging in the world of wine. For example, a consortium of

California-based wine organizations bestows “Green Medal” awards to California vineyards and wineries for honors such as exceptional leadership in “the three ‘Es’ of sustainability—Environmentally sound, socially Equitable, and Economically viable practices” as judged by a panel of experts in wine and sustainability (California Green Medal, 2022). Whether bottled water companies are operating in an environmentally and socially sustainable way could be a compelling basis for future water competitions.

Supplementary material. To view supplementary material for this article, please visit <https://doi.org/10.1017/jwe.2022.41>

Acknowledgments. The authors thank the managing editor, journal reviewers, and Jeff Bodington for sharing the 2019 California State Fair wine competition discussed herein. By way of disclosures, Berg was certified as a water sommelier through the Fine Water Academy during a sabbatical in 2020 and, subsequent to that, was one of the judges at the 2021 Fine Water Taste and Design Awards. As detailed in this paper, Mascha is affiliated with the Fine Water Society, Academy, and Taste and Design Awards. Capehart has no affiliation with those entities or events.

References

- Ashenfelter, O., and Quandt, R. (1999). Analyzing a wine tasting statistically. *Chance*, 12(3), 16–20.
- 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. (2017). Dimensions of expertise in wine evaluation. *Journal of Wine Economics*, 12(1), 59–83.
- Azoulay, A., Garzon, P., and Eisenberg, M. J. (2001). Comparison of the mineral content of tap water and bottled waters. *Journal of General Internal Medicine*, 16(3), 168–175.
- Biro, A. (2019). Reading a water menu: Bottled water and the cultivation of taste. *Journal of Consumer Culture*, 19(2), 231–251.
- Bitter, C. (2017). Wine competitions: Reevaluating the gold standard. *Journal of Wine Economics*, 12(4), 395–404.
- Bodington, J. (2020). Rate the raters: A note on wine judge consistency. *Journal of Wine Economics*, 15(4), 363–369.
- Burlingame, G. A., Dietrich, A. M., and Whelton, A. J. (2007). Understanding the basics of tap water taste. *American Water Works Association*, 99(5), 100–111.
- Cal Expo (2019). California Wine Competition 2019 Judging Photos. Available at <https://calexpostatefair.com/participate/competitions/california-commercial-wine/california-wine-competition-2019-judging-photos/> (accessed May 1, 2021).
- California Green Medal (2022). 2022 California Green Medal Winners. Available at <https://www.greenmedal.org/all-winners/2022-winners/> (accessed July 1, 2022).
- Capehart, K. W. (2015). Fine water: A hedonic pricing approach. *Journal of Wine Economics*, 10(2), 129–150.
- Capehart, K. W., and Berg, E. C. (2018). Fine water: A blind taste test. *Journal of Wine Economics*, 13(1), 20–40.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- Dietrich, A. M., and Burlingame, G. A. (2020). A review: The challenge, consensus, and confusion of describing odors and tastes in drinking water. *Science of the Total Environment*, 713, 1–11, doi: 10.1016/j.scitotenv.2019.135061.
- DiNardo, J., and Tobias, J. L. (2001). Nonparametric density and regression estimation. *Journal of Economic Perspectives*, 15(4), 11–28.
- Fine Water Academy (2022). Certified Water Sommelier. Available at <https://finewateracademy.talentlms.com/index> (accessed July 1, 2022).

- Fine Water Society (2016). First international fine water tasting competition in China with renowned water sommeliers [Press Release]. Available at <https://web.archive.org/web/20220614195049/http://finewaters.com/premium-bottled-water/finewaters-tasting-competitions/guangzhou-china-2016/264-press-release> (accessed June 14, 2022).
- Fine Water Society (2021). Jury panel 2021. Available at <https://finewaters.com/fine-water-society/taste-design-awards/bled-slovenia-2021/jury-panel>.
- Fulcher, J. (2017). Judging water at Berkeley Springs international water tasting competition. *Water Environment Federation Highlights*, April 7. Available at <http://news.wef.org/judging-water-at-berkeley-springs-international-water-tasting-competition/> (accessed April 7, 2017).
- Gergaud, O., Ginsburgh, V., and Moreno-Terreno, J. D. (2021). Wine ratings: Seeking a consensus among tasters via normalization, approval, and aggregation. *Journal of Wine Economics*, 16(3), 321–342.
- Gleick, P. H. (2010). *Bottled and Sold: The Story Behind Our Obsession with Bottled Water*. Washington, DC: Island Press.
- Hodgson, R. T. (2008). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.
- Hodgson, R. T. (2009). An analysis of the concordance among 13 U.S. wine competitions. *Journal of Wine Economics*, 4(1), 1–9.
- Hooks, C. (2013). Beyond Fiji and Perrier, with a water sommelier. *New York Times*, December 22. Available at <http://nyti.ms/JS7oJT> (accessed December 22, 2013).
- Hulkower, N. D. (2009). The judgment of Paris according to Borda. *Journal of Wine Research*, 20(3), 171–182.
- Marcussen, H., Holm, P., and Hansen, H. (2013). Composition, flavor, chemical foodsafety, and consumer preferences of bottled water. *Comprehensive Reviews in Food Science and Food Safety*, 12(4), 333–352.
- Mascha, M. (2022). *Fine Waters: A Connoisseur's Guide to the World of Premium Waters*. Available at <https://finewaters.com>.
- Oczkowski, E. (2017). The preferences and prejudices of Australian wine critics. *Journal of Wine Research*, 28(1), 56–67.
- Olkin, I., Lou, Y., Stokes, L., and Cao, J. (2015). Analyses of wine-tasting data: A tutorial. *Journal of Wine Economics*, 10(1), 4–30.
- Riese, Martin. (2022). How a water tasting event look like. *TikTok*, April 29. Available at <https://www.tiktok.com/@martinrieseofficial/video/7092105196326702379> (accessed April 29, 2022).
- Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, 13(1), 19–30.
- Taber, G. M. (2006). *The Judgment of Paris: California vs. France and the Historic 1976 Paris Tasting that Revolutionized Wine*. New York, NY: Scribner.
- Taber, G. M. (2012). The judgment of Princeton. *Journal of Wine Economics*, 7(2), 143–151.
- Tishgart, S. (2017). Ask a water sommelier. *Grub Street*, January 8. Available at <https://www.grubstreet.com/2017/01/whats-the-best-bottled-water.html> (accessed January 8, 2017).