**Supplementary Appendix**

*[FOR ONLINE PUBLICATION ONLY]*

The following are additional remarks and results related to those discussed in the main text. The data and code for this study can be found via the following link: <<https://osf.io/wfme6/>>.

**Clarifying note on the potential meaningfulness of a strong positive inter-rater correlation between wine judges’ ratings:**

In the main text of the paper in the context of discussing inter-rater correlations between wine judges’ ratings, we say that the judges’ collective judgment about relative wine quality *might* be meaningful if there is a strong positive correlation between their ratings. It should also be recognized that it *might not* be meaningful. An apparently strong correlation could have been purely coincidental, especially if there are only a few judges or wines. And even if the strong correlation was not pure coincidence, the consensus of the panel might not be useful information to anyone else. To the extent that wine is mostly an experience good, others might not perceive the same differences or share the same aesthetic standards. To the extent that wine is mostly a credence good whose quality is difficult to verify or even what Ekelund and Thornton (2019) call a “meta-credence” good whose quality is impossible to verify, others might not recognize the panel as the sort of authority that can provide quality assurances. Nevertheless, the collective judgment of the panel would seem to be more meaningful if there is a strong positive correlation rather than little or no correlation between the judges’ rating. A strong positive correlation (rather than no correlation or a negative correlation) means that the panel gave clear (rather than ambiguous or conflicting) guidance about the relative quality of the wines. The same could be said about waters or other items besides wines.

**Clarifying note on intra- (rather than inter-) rater reliability for the Taste Awards:**

As mentioned in the main text of the paper, there are no replicate samples for any given year of the Taste Awards. It is therefore not possible to examine *intra*-rater reliability for a given year of the Taste Awards. That is to say, it is not possible to examine whether a given judge rates the blind taste of a given water the same or similarly when repeatedly tasting it at one year of the Taste Awards.

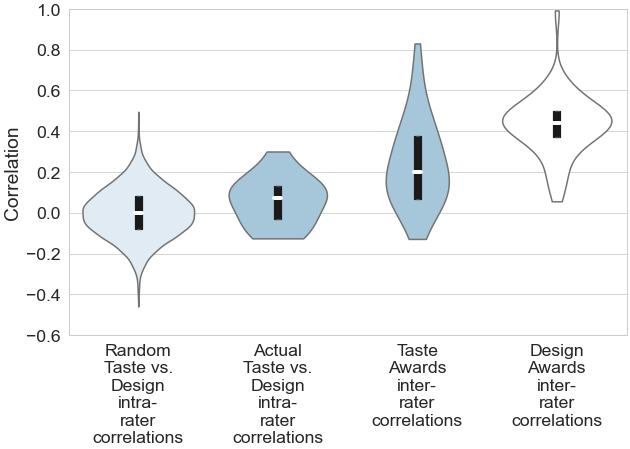
Intra-rater reliability *across years* of the Taste Awards could be studied because some judges and some waters reoccur over the years, although the resulting sample size would be small. Even with a larger sample size, some intra-rater un-reliability across years would not be unreasonable. A year or more would have passed since the last time they tasted the water.

**Additional finding on intra-rater correlation between Taste and Design Awards ratings:**

Although it is not possible to examine intra-rater reliability for any given year of the Taste Awards, we can examine whether there is any correlation between the rating that a judge gives to a bottled water in the Taste Awards and the rating that the same judge gives to the same bottled water’s container in the Design Awards. Correlations can be calculated for any bottled waters that compete in both the Taste and Design Awards in a given year.

To be clear, the question of whether there is a correlation between a judge’s ratings for the Taste and Design Awards in a given year is a different question from whether there would be any correlation between the judge’s ratings for multiple blind tastings of the same water in a relatively short period of time. Yet it is potentially of interest to know whether a water that tastes good in a blind tasting comes in attractive bottles. That is of particular interest because, given that the Design Awards occur immediately before the Taste Awards, and also given that most (although not all) waters compete in both parts of the competition, a strong positive intra-rater correlation between judges’ ratings for the Taste and Design Awards might make us worry that the judges were able to accurately guess which waters were which and, in turn, they may have been influenced by branding or other factors that are supposed to be masked by a blind tasting.

The following figure, which we explain below, shows that the correlation between a given judge’s rating of a water for the Taste Awards and the same judge’s rating of the water’s bottle for the Design Awards is generally about the same as what would be expected by random chance alone.



Just like the figure in the main text, the above-given figure shows a violin plot of the inter-rater Pearson correlation coefficients for the Taste Awards (labeled “Taste Awards” in the main text and “Taste Awards inter-rater correlations” here) and a violin plot of similar correlations for the Design Awards (labeled “Design Awards” in the main text and “Design Awards inter-rater correlation” here).

Unlike the figure in the main text, the above-given figure also shows a violin plot of intra-rater Pearson correlation coefficients for a given judge’s ratings of a water for the Taste Awards and the same judge’s rating of the water’s bottle for the Design Awards (labeled “Actual Taste vs. Design intra-rater” correlations). Those correlations were calculated across all the waters that a judge judged across their years in the competition. Some judges have judged for as little as one year, while others have judged for all five years (as detailed in the main text).

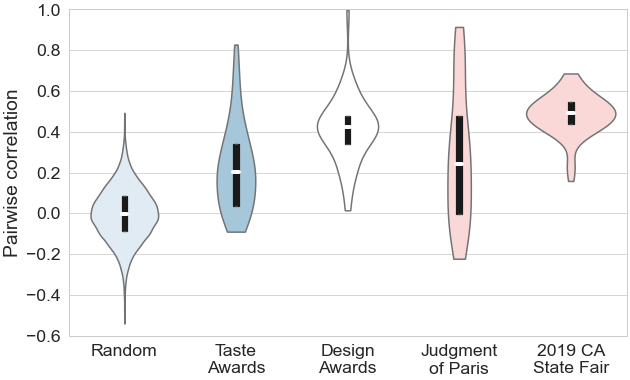
The distribution of intra-rater correlations has a range of −0.13 to 0.30, the interquartile range is −0.03 to 0.13, the median is 0.07, and the average is 0.06.

The above-given figure also shows the distribution of intra-rater correlations that would have been observed if each judge was randomly assigning ratings as part of both the Taste and Design Awards (labeled “Random Taste vs. Design intra-rater correlations”). As can be seen in the figure, the actual and randomly generated distributions are fairly similar. The PS (again, probability of superiority) for the actual distribution over the randomly generated one is 63% (p-value=0.08, again based on a Mann-Whitney U test with the null hypothesis that the PS is 50%).

We therefore conclude that the intra-rater correlation for a judge’s rating of a bottled water for the Taste Awards and the same judge’s rating of the bottled water’s container for the Design Awards is generally no better than what would be expected by chance.

**Additional finding: The results of the study are similar if the Spearman (rather than Pearson) correlation coefficient is used.**

As stated in the main text, the results presented in this study are almost identical if we use the Spearman rank-order correlation coefficient rather than the Pearson product-moment correlation coefficient. That can be seen in the following figure, which is constructed in the exact same way as the figure in the main text, except it uses the Spearman (rather than Pearson) correlation coefficient.



*Same as Figure 1 in the main text, except using Spearman correlation coefficient*

There are no notable differences between that figure and the one in the main text. If there had been any notable differences, then the results using the Spearman rather than Pearson correlation coefficient would be (at least arguably) the more appropriate ones. Ashenfelter and Quandt (1999), Gergaud, Ginsburg, and Moreno-Ternero (2021), and others have argued in favor of using ordinal rankings when evaluating wines and wine judging (as already stated in the main text).

**Additional finding: For the 13 previous blind-wine-tasting studies identified by Ashton (2012) and Oczkowski (2017), the inter-rater reliability of wine experts’ ratings varies.**

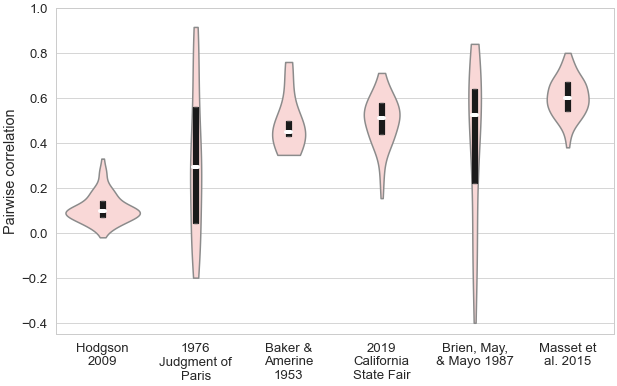
As noted in the main text, Ashton (2012) identified nine studies that reported pairwise Pearson correlation coefficients over which a simple average could be taken for different wine experts rating wines based on blind tasting (with some exceptions flagged in the main text and in the table below). While discussing Ashton’s (2012) findings, Oczkowski (2017) identified four more studies that did the same. The following table provides details on those 13 studies.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Study** | **Judges were experts at a wine competition?** | **Average pairwise Pearson correlation coefficient** |
| 1 | Baker and Amerine’s (1953) study of five lab subjects blind tasting and rating 13 red wines (specifically, Cabernets). The subjects were “highly trained tasters who had an extensive acquaintance with both Cabernet and Riesling wines. The tasters had had experience in wine tasting for periods of from 5 to 15 years; 3 of them had participated in the wine judgings at the California State Fair” (p. 386). | No, but subjects were experts completing a lab experiment similar to a wine competition | 0.39 average (per Ashton, 2012) |
| 2 | Same as above, except 17 white wines (specifically, Rieslings) | “ “ | 0.58 average (per Ashton, 2012) |
| 3 | Brien, May, and Mayo’s (1987) Occasion 1 of Study 2 in which six judges rated 24 Shiraz wines | No, but subjects were experts completing a lab experiment similar to a wine competition | 0.45 average (per Ashton, 2012) |
| 4 | Brien, May, and Mayo’s (1987) Occasion 2 of Study 2, in which the same six judges rated the same 24 Shiraz wines later on the same day | “ “ | 0.37 average (per Ashton, 2012) |
| 5 | Cicchetti’s (2004) analysis of red wines at Judgment of Paris | Yes | Not applicable; an intraclass (not Pearson) correlation coefficient of 0.22 was reported |
| 6 | Cicchetti’s analysis of white wines at Judgment of Paris | Yes | Not applicable; an intraclass (not Pearson) correlation coefficient of 0.36 was reported. Ashton (2012) cites an unpublished paper by Cicchetti for that statistic. The same statistic was reported in: Cicchetti and Cicchetti (2008). |
| 7 | Ashton’s (2011) analysis of red wines at Judgment of Paris | Yes | 0.16 average (per Ashton, 2012) |
| 8 | Ashton’s (2011) analysis of white wines at Judgment of Paris | Yes | 0.44 average (per Ashton, 2012) |
| 9 | Hodgson’s (2009) analysis of awards from 13 U.S. wine competitions | No, but “judges” were different wine competitions | 0.11 average (per Ashton, 2012) |
| 10 | Ashton’s (2013) study of six wine critics’ (including, for example, Robert Parker’s) published ratings of Bordeaux wines | No, wine critic’s published ratings. Also, the tastings might not have actually been blind because they were Bordeaux *en primeur* tastings. Ashton (2013) notes, as others have, that “Bordeaux *en primeur* tastings are generally not blind” (p. 233). | 0.60 average (per Oczkowski, 2017) |
| 11 | Masset et al.’s (2015) study of 12 wine critics’ (including, for example, Robert Parker’s) published ratings of Bordeaux wines | No, wine critics’ published ratings for Bordeaux *en primeur* tastings | 0.604 average (per Oczkowski, 2017) |
| 12 | Cardebat and Paroissien’s (2015) 16 wine critics’ (including, for example, Robert Parker’s) published ratings of Bordeaux wines. | No, wine critics’ published ratings for Bordeaux *en primeur* tastings | 0.59 average (per Oczkowski, 2017) |
| 13 | Stuen, Miller, and Stone’s (2015) study of the ratings of 4 wine critics (*Wine Spectator, The Wine Advocate, Wine Enthusiast, and International Wine Cellar*) published ratings of Washington and California wines | No, wine critics’ published ratings | 0.463 average, per Oczkowski (2017), where by our calculations, 0.4683 was the average across the 45 wines that all 4 critics rated |

As noted in the main text and detailed in the table above, only two of the nine wine-tasting studies identified by Ashton (2012) and none of the four studies identified by Oczkowski (2017) were actually studies of the pairwise Pearson correlation coefficients for judges’ ratings at a wine competition. Those two studies were Ashton’s (2011) analyses of the white wine flight and, separately, the red wine flight at the 1976 Judgment of Paris.

In terms of the other 11 studies: Four of the ones identified by Ashton (2012) and none of the ones identified by Oczkowski (2017) were lab experiments set up similarly to wine competitions. The remaining studies either used an intraclass correlation coefficient (in the case of two studies identified by Ashton 2012, which were analyses of the Judgment of Paris), were Hodgon’s (2009) study of what if any awards were bestowed by different wine competitions, or were studies of wine critics’ published ratings (in the case of all four studies identified by Oczkowski 2017, three of which studies of Bordeaux *en primeur* tastings).

Although some of those studies do not seem entirely comparable to each other, if we treated them as directly comparable, then the studies suggest at least somewhat different conclusions about inter-rater reliability for wine experts’ ratings. We can draw the following figure in order to illustrate that.



Like the figure in the main text, the above-given figure shows a violin plot of the pairwise Pearson correlation coefficients for the Judgment of Paris (where the correlations were calculated across all 20 wines for each pair of the 11 judges) and the 2019 California State Fair (where the correlations were calculated across all the wines at once for each pair of judges that tasted the same wines).

Unlike the figure in the main text, the above-given figure also shows the following. On the far left and far right of the figure are violin plots of the pairwise Pearson correlation coefficients from Hodgson’s (2009) and Masset *et al.*’s (2013) studies, respectively. Those two studies are the ones with the lowest and highest average pairwise Pearson correlation coefficients among the 13 studies identified by Ashton (2012) and Oczkowski (2017). Hodgon’s (2009) study of the awards bestowed by different wine competitions had the lowest average at 0.11, whereas Masset *et al*.’s (2013) study of the ratings assigned by wine critics to Bordeaux *en primeur* tastings had the highest average at 0.604.

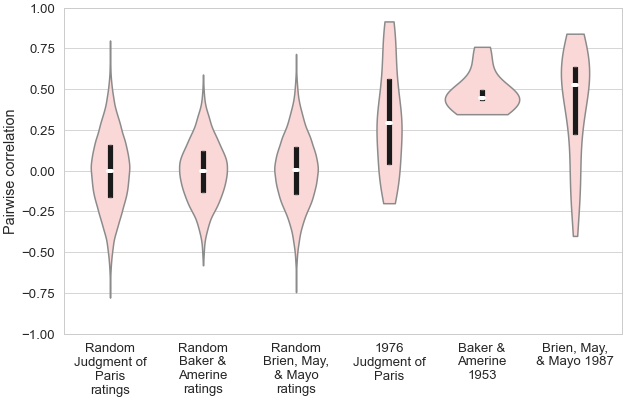
The figure also shows a violin plot of the pairwise Pearson correlation coefficients from Baker and Amerine’s (1953) lab experiment. They reported judge-level scores. We used those scores to calculate correlations for each pair of the five lab subjects over all the wines blind tasted and rated (which were 17 Rieslings and 13 Cabernets, or 30 wines in total).

Finally, the figure shows a violin plot of the pairwise Pearson correlation coefficients from Brien, May, and Mayo’s (1987) lab experiment. They did not report judge-level ratings, but they did report two sets of inter-rater pairwise correlations for when six lab subjects blind tasted and rated 24 Shiraz wines on separate occasions. We used those reported correlations.

The violin plots are arranged by their median pairwise correlation in ascending order from left to right. As can be seen in the figure, there is considerable variation in the medians of those distributions of pairwise correlations, their interquartile ranges, and other aspects of the distributions. There is variation between the Judgment of Paris and 2019 California State commercial wine competition, in particular, which are the two wine competitions we focus on in the main text.

**Additional finding: The inter-rater reliability of the wine judges at the 1976 Judgment of Paris is generally better than random chance based on Bodington’s (2020) method. The same is also true for the wine experts in the two lab experiments identified by Ashton (2012).**

Like the figure above, the figure below shows violin plots for the distribution of pairwise Pearson correlation coefficients for the judges’ ratings at the Judgment of Paris (again labeled “1976 Judgment of Paris”) and the two lab experiments identified by Ashton (2012) that were set up similarly to wine competitions (labeled “Baker & Amerine 1953” and “Brien, May, & Mayo 1987”). The figure also shows the distribution of pairwise Pearson correlation coefficients that would have been observed if the experts were randomly assigning ratings to the wines they tasted, as discussed further below.



To generate the distribution of pairwise Pearson correlation coefficients that would be observed if the experts were randomly assigning ratings, we followed Bodington (2020) and started with information on the setup of a competition or competition-like lab experiment (in terms of the number of occasions on which wines were tasted, the number of wines tasted on a given occasion, the number of experts rating those wines, and the ratings the experts could have possibly assigned to a wine). For the Judgment of Paris, there were 11 judges rating 20 wines on a zero to 20 point scale. For Baker and Amerine’s (1953) lab experiment, five experts rated 30 wines on a 100-point scale. For the Brien, May, and Mayo’s (1987) lab experiment, six experts rated 24 wines on a 20-point scale once and then they did the same thing again on a second occasion. Using that information, we then assumed each expert randomly assigned a rating to each wine by drawing from a uniform distribution over the possible ratings. The Pearson correlation coefficients for those ratings could then, in turn, be calculated for each possible pair of judges. Violin plots of the distribution of pairwise correlations based on those randomly generated ratings are labeled as “Random Judgment of Paris ratings,” “Random Baker & Amerine 1953 ratings,” and “Random Brien, May, & Mayo 1987 ratings” in the figure above. Those randomly generated distributions all look fairly similar because their respective setups were all fairly similar in terms of the number of wines tasted.

The randomly generated distributions are all noticeably different from the distributions of pairwise correlations that were actually observed. The following table quantifies those differences by reporting the PS (again, “probability of superiority”) for an observed distribution of pairwise correlations over the corresponding randomly generated one. The p-value from a Mann-Whitey U test is also reported; the null hypothesis for that test is that the PS is 50%.

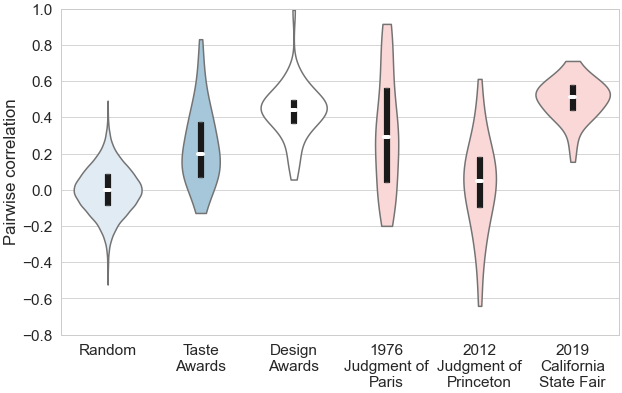
|  |  |  |
| --- | --- | --- |
| **Competition or experiment** | **PS for observed distribution of pairwise correlations over randomly generated distribution** | ***p*-value** |
| 1976 Judgment of Paris | 77% | <0.01 |
| Baker and Amerine (1953) | 99% | <0.01 |
| Brien, May, and Mayo (1987) | 83% | <0.01 |

*Note:* This table reports, for each competition or lab experiment discussed above, the PS (probability of superiority) for the distribution of the pairwise Pearson correlation coefficients that were actually observed, on the one hand, over the distribution of pairwise Pearson correlation coefficients that would have been observed if the experts were randomly assigning ratings. The *p*-value is for a Mann-Whitey U test of whether the PS is different from 50%.

We therefore conclude based on Bodington’s (2020) method that, especially for the lab experiments and somewhat less so for the Judgment of Paris, the pairwise correlations between the experts’ ratings are generally higher than what would be expected by random chance alone.

**Additional finding: The inter-rater reliability of the wine judges at the 2012 Judgment of Princeton is somewhat worse than that of the wine judges at the 1976 Judgment of Paris and no better than random chance.**

The following figure is exactly like the one in the main text, except we have added a new part labeled “2012 Judgment of Princeton.” That part is a violin plot of the pairwise Pearson correlation coefficients for the nine judges who blind tasted and rated 20 wines on a zero to 20 point scale as part of the 2012 Judgment of Princeton. That “judgment” was designed to be similar to the famous one in Paris, except wines from New Jersey (rather than California) and France were tasted. At that event, the judges tasted white wines and red wines in two flights; but as elsewhere, we calculated the correlation coefficients across all wines at once.



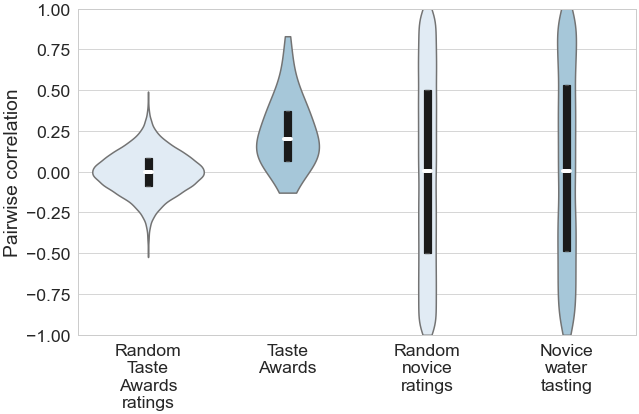
The median pairwise correlation between the judges’ ratings for the Judgment of Princeton was about 0.05, the interquartile range was from −0.10 to 0.18, and the range was from −0.64 to 0.61. The average was 0.03. The PS for the distribution of the Judgment of Paris correlations over the distribution of the Judgment of Princeton correlations is 74% (p-value < 0.01). The PS for the distribution of the Taste Awards correlations over the Judgment of Princeton correlations is similar at 73% (*p*-value < 0.01).

Although it is not shown in the figure immediately above, if the judges at the Judgment of Princeton had been randomly assigning ratings, then the distribution of pairwise Pearson correlation coefficients we would expect to observe would look essentially identical to the violin plot labeled “Random Judgment of Paris ratings” in the figure given further above (because the Judgements of Paris and Princeton were set up similarly). The distribution of correlations that were actually observed at the Judgment of Princeton would be largely indistinguishable from the distribution of correlation that would have been expected if judges were randomly assigning ratings. The PS for the former distribution over the latter would be only 54% and not statistically significantly different from 50% (*p*-value = 0.39).

We therefore conclude that the inter-rater reliability of the wine judges at the 2012 Judgment of Princeton is somewhat worse than that of the wine judges at the 1976 Judgment of Paris, somewhat worse than that of the water judges at the Taste Awards, and no better than random chance. Such a conclusion is not necessarily an indictment of the expertise of the judges at the Judgment of Princeton; perhaps the differences within and between the wines from New Jersey and France were not distinguishable enough to yield any strong (positive or negative) correlations between the judges’ ratings for those wines.

**Additional finding: The inter-rater reliability of water novices blind tasting and rating bottled waters is no better than chance and worse than that of the water experts at the Taste Awards.**

The following figure, which we describe in detail further below, compares the inter-rater reliability for water novices and experts blind tasting and rating bottled waters.



Just like the figure in the main text, part of the above-given figure shows a violin plot of the pairwise Pearson correlation coefficients for the judges’ ratings from the Taste Awards. That part of the figure was labeled “Taste Awards” in the main text and labeled the same here. The figure also shows a violin plot of what the pairwise correlation coefficients would have been if the judges in the Taste Awards assigned ratings randomly. That was labeled as “Random” in the similar figure in the main text. We have relabeled it as “Random Taste Awards ratings” here for additional clarity.

A new part of the figure, which is labeled “Novice water tasting,” is a violin plot of the pairwise Pearson correlation coefficients for the roughly 100 students who blind tasted and rated four bottled waters, as well as a local tap water, as part of Capehart and Berg’s (2018) study. Pearson correlation coefficients for students’ ratings of the four bottled waters were calculated across all possible pairs of students, except one student who gave the same rating to each water. (The Pearson correlation coefficient between that one student’s non-varying ratings and any other student’s ratings would be undefined.) We have drawn that part of the figure in order to compare the inter-rater reliability of those students who blind tasted and rated bottled waters, on the one hand, to the inter-rater reliability of the Taste Award judges doing the same, on the other hand.

It should be recognized that the students’ ratings from the Capehart-Berg study and the judges’ ratings from the Taste Awards are not necessarily comparable. In terms of the bottled waters: The Capehart-Berg study selected bottled waters that were identified as fine waters by water experts, but none of the four brands of bottled water students tasted (Speyside Glenlivet, Acqua Panna, Fiji, Hildon) have competed in the Taste Awards (or the Design Awards, for that matter). The bottled waters that students tasted were also all still waters with no sparkling ones. Those bottled waters also had a narrower range of TDS levels (from 58 to 312 mg/L) than the waters competing in the Taste Awards (where, for example, the still waters that won Gold medals at the 2021 Taste Awards for the “super low” and “high” minerality categories had TDS levels of 44 and 829 mg/L, respectively). In terms of the tasters: The students should not be considered water experts and should instead be considered novices, especially compared to the judges at the Fine Water Taste and Design Awards. Another difference is that the Taste Awards involve a small number of judges rating a large number of waters, whereas the Capehart-Berg study involved a large number of students rating a small number of waters. The comparison was nevertheless potentially interesting, so we compared them and are reporting the results here.

As seen in the figure above, the median pairwise correlation between the students’ ratings was essentially zero, the interquartile range was from −0.48 to 0.53, and the range was from negative to positive unity. The average was a mere 0.02. In addition to having a median and mean of essentially zero, the other aspects of the distribution of the students’ pairwise correlations are strikingly similar to what it would look like if the students randomly assigned ratings. Indeed, if we assume that the roughly 100 students drew their ratings for the four waters from a uniform distribution over the 1 to 14-point scale on which they rated the waters, then the median and mean of the pairwise correlations would both be essentially zero, the interquartile range would be from negative to positive 0.50, and the range would be from negative to positive unity. A violin plot of the pairwise correlation coefficients for those simulated ratings is shown as the “Random novice ratings” part of the figure above.

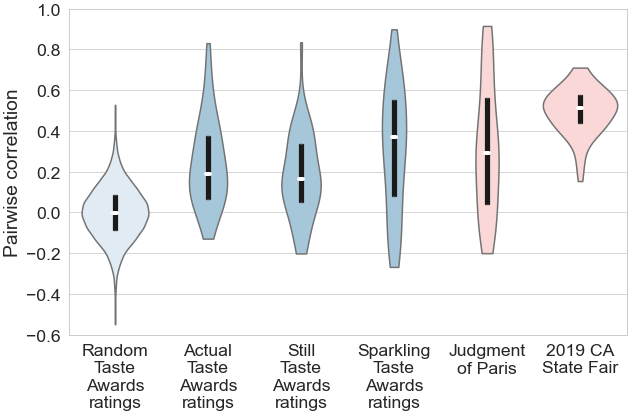
We therefore conclude that, unlike the water experts’ ratings from the Taste Awards, the students’ ratings from the Capehart-Berg study are almost entirely indistinguishable from random chance. That finding suggests, as we point out in the main text of the paper, that water experts may have some expertise that novices lack; the novices may lack the ability to identify differences between waters and/or they may lack shared aesthetic standards for translating identified differences into ratings. That said, the novices may have had a more difficult task than the experts. Whereas the novices judged waters that were all still and had a relatively narrow range of TDS levels, the experts judged waters that varied in their carbonation levels and had a wider range of TDS levels.

**Additional finding: The inter-rater reliability for the sparkling waters at the Taste Awards is somewhat higher than that for the still waters.**

As discussed in the main text, we calculate correlation coefficients over all the waters tasted at the Taste Awards in a given year. Reasons for doing that include that the experts taste the waters one at a time in a continuous stream, some of the competition categories have only a small number of waters, and although we should perhaps not question the expertise of those who organized the waters into categories, the categorization may be somewhat arbitrary.

We can try to examine inter-rater reliability over different groups of waters separately rather than all waters at once. At least one way to group waters while still ensuring fairly large sample sizes is by grouping waters according to whether they are still or sparkling. The figure below is just like the figure in the main text, except we also show the distribution of pairwise Pearson correlation coefficients if the correlations were calculated over the still waters (labeled “Still Taste Awards ratings”) and, separately, over the sparkling waters (labeled “Sparkling Taste Awards ratings”). When making those calculations, we ignored the relatively small number of waters that competed in either the curated or exotic water categories because it is not clear from those category names alone whether the waters were still or sparkling. For all the other categories, it is clear whether the waters were still or sparkling (either “naturally” or otherwise). Calculated in that way, there have been 202 still waters rated across the five years of the competition and 127 sparkling ones (where those numbers include any unique waters that competed in multiple years). The sample sizes for the still and sparkling waters are somewhat dissimilar, therefore, but each yields a fairly large number of ratings and pairwise correlations over those ratings.

Violin plots of the pairwise correlations over the still and sparkling water ratings are shown as part of the figure below, which is otherwise similar to the figure in the main text. For the distribution of pairwise correlations over the *still* waters, the range is −0.22 to 0.84, the interquartile range is 0.06 to 0.33, the median is 0.17, and the average is 0.18. For the distribution of pairwise correlations over the *sparkling* waters, the range is −0.27 to 0.90, the interquartile range is 0.08 to 0.55, the median is 0.37, and the average is 0.31. The PS for the latter distribution over the former is 64% (p-value=0.02). The inter-rater reliability for sparkling waters is therefore perhaps somewhat higher than that for still waters.

****

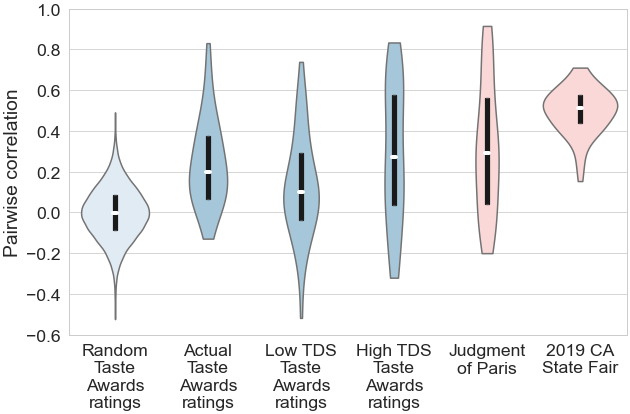
**Additional finding: The inter-rater reliability for relatively high TDS waters at the Taste Awards is somewhat higher than that for lower TDS waters.**

Another manner in which waters at the Taste Awards can be grouped while still maintaining fairly large sample sizes is by whether they have a relatively high or low TDS (again, total dissolved solids) level. As mentioned in the main text, the Fine Water competition groups still waters and, separately, sparkling waters into categories based on their TDS levels. For still waters, there are categories for “super low minerality,” “low minerality,” “medium minerality,” and “high minerality.” Similarly, for sparkling waters, there are categories for low, medium, and high minerality.

There were also categories for “added” and “naturally” sparkling waters, but those category names alone do not reveal TDS levels, so ignore those categories in the calculations that follow. We also ignore the curated and exotic categories in our calculations because those category names do not reveal TDS levels, either. We leave it as a direction for future research to merge the dataset we analyze here with a dataset that would detail the TDS levels and other objective properties of each water that has competed in the Taste Awards.

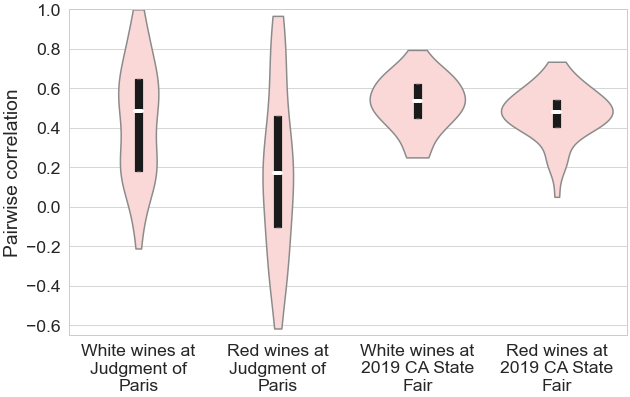
If we combine the still and sparkling waters in the “super low” and “low” minerality categories, on the one hand, and the still and sparkling waters in the “medium” or “high” minerality categories, on the other hand, then there have been 157 waters with relatively low TDS levels rated across the five years of the competition and 86 waters with relatively higher TDS levels rated. Much like when the sample is divided into still and sparkling waters, those sample sizes are fairly large but somewhat dissimilar.

Violin plots of the pairwise correlations over the ratings of waters with relatively low and high TDS levels are shown as part of the figure below, which is otherwise similar to the figure above and the figure in the main text. For the distribution of pairwise correlations over the waters with relatively *low* TDS levels, the range is −0.52 to 0.74, the interquartile range is −0.04 to 0.29, the median is 0.10, and the average is 0.13. For the distribution of pairwise correlations over waters with relatively *high* TDS levels, the range is −0.32 to 0.83, the interquartile range is 0.04 to 0.58, the median is 0.27, and the average is 0.29. The PS for the latter distribution over the former is coincidentally almost identical to the differences between the still and sparkling waters (i.e., 64% with a p-value=0.02). The inter-rater reliability for high TDS waters is therefore perhaps somewhat higher than that for low TDS waters.



**Additional finding: The inter-rater reliability for red wines at wine competitions is somewhat higher than that for white wines.**

Ashton’s (2012) study suggests that the inter-rater reliability for red wines may be higher than that for white wines. As seen in the figure below, the distribution of pairwise Pearson correlation coefficients calculated over just the white wine flight at the Judgment of Paris is generally somewhat higher than the distribution of correlations calculated over just the red wine flight. The PS for former over the latter is 68% (*p*-value<0.01).



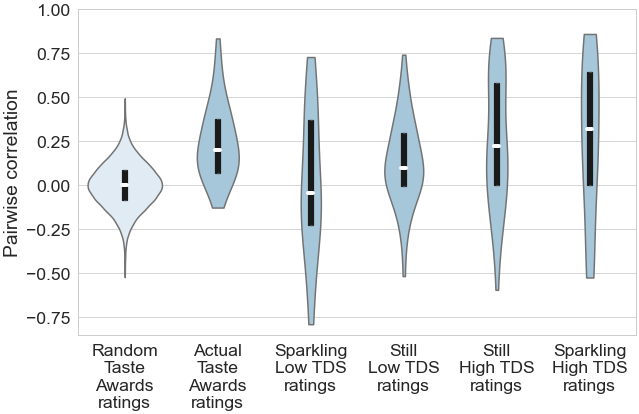
The difference is smaller for the 2019 California State Fair, but the above-given figure also shows that the distribution of pairwise Pearson correlation coefficients calculated over just the white wines at that competition is somewhat higher than the distribution of correlations calculated over the red wines (where we ignored any sparkling red wines because they were lumped together with rosé and blush sparkling wines). The PS for former over the latter is 61% (*p*-value=0.04).

It can be noted that the difference in inter-rater reliability between still and sparkling waters at the Taste Awards, as well as between waters with relatively low and high TDS levels, are similar in size to the difference in inter-rater reliability between white and red wines at wine competitions. By noting that, we do not mean to strain the analogy between water and wine or suggest that red wine is to still or low TDS wine as white wine is to sparkling or high TDS water. Similar differences may also exist for other ways in which waters and wines can be grouped together.

**Additional finding: The inter-rater reliability for sparkling waters with relatively high TDS waters may be somewhat higher than that for still or lower TDS waters.**

As discussed above, inter-rater reliability seems to be somewhat higher for sparkling waters compared to still ones, as well as for waters with relatively high TDS levels compared to waters with lower TDS levels. It is therefore potentially of interest to examine whether inter-rater reliability varies between, say, sparkling waters with relatively high TDS levels and still waters with relatively low TDS levels. The sample sizes will obviously be even smaller than the above-mentioned ones with even narrower groupings, but for completeness, we can examine it.

The figure given below shows violin plots of the pairwise correlations over those more narrowly defined groups of still or sparkling waters with relatively low or high TDS levels. The number of waters underlying some of those violin plots are quite small. In particular, due to the way in which we defined the groupings (ignoring the added and natural carbonation categories whose category names alone do not reveal TDS levels) only 21 waters have been tasted that we classify as both sparkling and also having relatively high TDS levels. Only 29 have been tasted that we classify as sparkling and having relatively low TDS levels. For the still waters, the sample sizes are larger; there have been 128 and 65 with relatively low and high TDS levels, respectively.



Due to the small sample sizes for some groupings, any conclusions drawn from the above-given figure should be treated as tentative. With that caveat in mind, the conclusions we can draw from the figure are mostly consistent with the results discussed further above. Inter-rater reliability again seems to be somewhat higher for sparkling waters with relatively high TDS levels, with the exception that the pairwise correlations for sparkling waters with relatively low TDS levels are perhaps somewhat lower than the correlations for still waters with relatively low TDS levels.

**References not already cited in the main text**

**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.

**Baker, G.A., and Amerine, M.A.** (1953). Organoleptic ratings of wines estimated from analytical data. *Food Research*, **18**(1–6), 381–389.

**Brien, C.J., May, P. and Mayo, O.** (1987). Analysis of judge performance in wine‐quality evaluations. *Journal of Food Science*, **52**(5), 1273–1279.

**Cardebat, J-M., and Paroissien, E.** (2015). Standardizing expert wine scores: An application for Bordeaux *en primeur*. *Journal of Wine Economics*, **10**(1), 329–348.

**Cicchetti, D.V.** (2004). Who won the 1976 blind tasting of French Bordeaux and U.S. cabernets? Parametrics to the rescue. *Journal of Wine Research*, **15**(3), 211–220.

**Cicchetti, D.V., and Cicchetti, A.** (2008). The balancing act in consistent wine tasting and wine appreciation: Part II: consistency in wine tasting and appreciation: an empirical-objective perspective. *Journal of Wine Research*, **19**(3), 185–191.

**Ekelund, R.B., and Thornton, M.** (2019). Extreme credence and imaginary goods. *Atlantic Economic Journal*, **47**(3), 361–371.

**Masset, P., Weisskopf, J-P., and Cossutta, M.** (2015). Wine tasters, ratings and *en primeur* prices. *Journal of Wine Economics*, **10**(1), 75–107.

**Stuen, E. T., Miller, J. R., and Stone, R. W.** (2015). An analysis of wine critic consensus: A study of Washington and California wines. *Journal of Wine Economics*, **10**(1), 47–61.