Tbilisi 2022 Abstract Submission

Title
Sources of Bias Remain in Blind Wine Ratings
'stuff that's not in the glass

I want to submit an abstract for:
Conference Presentation

Corresponding Author
Jeff Bodington

E-Mail Corresponding Author
jcb@bodingtonandcompany.com

Affiliation
Bodington & Company

Keywords
wine, ratings, bias

Research Question
What affects blind wine ratings other than the actual wine in the glass?

Methods
Literature review, WSET training and tasting experience

Results
Even blind wine ratings are influenced by factors that are not in the glass including stochastic, anchoring, expectation, serial position, and commercial biases.

Abstract
I. Introduction

The ratings that judges and critics assign to wines in newsletters, blogs, magazines, and in local to international competitions, affect consumers' decisions and the economics of the wine industry. While many of those ratings are assigned by tasters who are "blind" to the wines' prices and vintners, several sources of potential bias remain. Even when a wine is assessed blind, the rating assigned may be influenced by factors that are not in the glass. Ratings may be affected by stochastic, anchoring, expectation, serial position, and commercial biases.

II. The Maelstrom About a Rating Observed

Judges assign ratings to wines that are within a bounded set of scores, or an ordered set of categories, or a set of ranks. Examples of scores include the 50- to 100-point scales used by Wine Advocate and Wine Spectator, U.C. Davis' zero to 20-point scale, Jacis Robinson's 12- to 20-point scale, and the zero- to 100-point scale prescribed by the International Organization of Wine and Vine (OIV). Examples of categories include the Wine & Spirit Education Trust (WSET) six categories of quality (faulty, poor, acceptable, good, very good, and outstanding) and the California State Fair Commercial Wine Competition's (CSF) six or ten medals. Further, some systems are forced rankings. If there are six wines in a flight, a judge must rank all six in order of relative preference. Liquid Assets and San Francisco FOG are examples of tasting groups who employ that approach. See reviews and comparisons of rating systems in Cicchetti & Cicchetti (2014), Kliparchuck (2013), and Veseth (2008).

Although judges and critics focus on the wine in the glass, much research shows that the ratings that tasters assign are affected by other factors. Some of the factors are supported by literature that is cited below. Other factors described below are reported as anecdotal and, when no literature is cited, those other factors are intended as hypotheses that remain to be tested.

II.1 Stochastic Ratings

First, although wine ratings are not merely random, evidence that ratings are stochastic is abundant in the wine-related academic literature and trade press. Bodington (2017, 2020) summarizes and cites four experiments with blind replicates, more than 20 other evaluations that find uncertainty in ratings, and two texts that explain the neurological, physiological, and psychological reasons for variance in the ratings that the same judge assigns to the same wine. The stochastic nature of wine ratings is not unique. Kahneman et al. (2021, p. 80-86, 215-258) describe variance in wine ratings and in many other areas of human judgement including physicians' diagnoses, forensic experts' fingerprint identifications, and judges' sentencings of criminals.

The wine-related literature cited above supports several findings. A rating observed is one draw from a latent distribution, it is one instance of in some cases many potential instances. Ratings are heteroscedastic, so the distribution of ratings on a wine is wine- and judge-specific and different judges' ratings on the same wine are not identically distributed (ID). Some judges assign ratings more consistently than others, and some wines are more difficult to rate consistently than others. Research attempts to predict ratings from physiochemical properties have struggled to obtain statistical significance. Experiments with blind replicates show that, on average, the standard deviation of the rating that the same judge assigns to the same wine within a flight is approximately 1.3 out of 10 potential rating categories. And while some judges independently assign ratings that correlate well with each other, about 10% of CSF judges assign ratings that are indistinguishable from random assignments.

II.2 Anchoring, Expectation, Serial Position, and Commercial Biases

The score-based rating systems noted above also assign categories of quality or award to score thresholds and ranges. For example, the OIV system sets score thresholds for Bronze, Silver, and Gold medals at scores of 80, 85 and 90 respectively. With a sample of 8,400 ratings, Bodington and Malfeito (2018) showed spikes in the frequencies of scores assigned just below those thresholds. Thus, some judges appear to anchor at OIV's category thresholds.

In addition to anchoring scores to category thresholds, there is anecdotal evidence of sequential anchoring. In a taste-and-score sequential protocol, a judge may assign a rating to the first wine and then rate the remaining wines "around" that anchor.

Much research shows that judges' expectations affect the ratings they assign. Ashton (2014) found that judges assigned higher ratings to wines from New Jersey when told the wines were from California and lower ratings to wines from California when told the wines were from New Jersey. On that evidence, regardless of actual quality, an expectation of good quality may lead to a central tendency in ratings within whatever range of scores or categories indicates good quality. In addition, information provided about wines may alter expectations and ratings. For example, the pre-printed forms provided to CSF judges list the grape variety, vintage, alcohol by volume and residual sugar of a wine next to spaces where the judge writes in a comment and then a rating. Whether or not such judgments should be represented as "blind" is open to debate. Whether or not having that information affects judges' ratings remains to be tested.

Profit-seeking may cause, in some cases and when ratings are not assigned blind, a commercial bias in ratings. Gray (2014, 2022) reports that wine ratings have been leaked to wine sellers in advance of publication, retailers with non-public information have cornered markets for some wines, and that ratings have been influenced by vintners' payments for advertising, sponsorships of tables at events, and other "pay to play" actions. Teeter (2014) and Siegel (2019) assert that wine critics' acceptances of free samples, dinners, and trips affect both what they write about a wine and the rating that they assign to a wine. Gregutt (2022) asserts that some critics inflate scores to get publicity for themselves. Those practices decouple wine ratings from the intrinsic qualities of wine in a glass and they may also affect research concerning the relationship between wine ratings and prices. Those possibilities are mentioned here for completeness and the author is not aware of any published assessment of the extent of such practices.

III. Conclusion

Even when wines are assessed blind to price and vintner, much evidence shows that, although they are not merely random, the ratings that critics and judges assign may be influences by stochastic, anchoring, expectations, serial position, and commercial biases.

File Upload (PDF only)

- SourcesOfBias-Bodington-3-9-22-abstract.pdf

Consent

✔ I agree to the privacy policy.

You find the link to our privacy policy at the bottom of the page. In the privacy policy you find a link for exporting and/or erasing your personal data stored in our database.

Sources of Bias in Wine Ratings *

’stuff that’s not in the glass

Jeff Bodington [a]

Abstract Submission for AAWE Tibisi 2022

I. Introduction

The ratings that judges and critics assign to wines in newsletters, blogs, magazines, and in local to international competitions affect consumers’ decisions and the economics of the wine industry. While each rating is assigned to a wine in a glass, much research shows that ratings are influenced by factors that are not in the glass. Ratings are affected by stochastic, anchoring, expectation, serial position, and commercial biases.

II. The Maelstrom About a Rating Observed

Judges assign ratings to wines that are within a bounded set of scores, or an ordered set of categories, or a set of ranks. Examples of scores include the 50- to 100-point scales used by *Wine Advocate* and *Wine Spectator*, U.C. Davis’ zero to 20-point scale, Jacis Robinson’s 12- to 20-point scale, and the zero- to 100-point scale prescribed by the International Organization of Wine and Vine (OIV). Examples of categories include the Wine & Spirit Education Trust (WSET) six categories of quality (faulty, poor, acceptable, good, very good, and outstanding) and the California State Fair Commercial Wine Competition’s (CSF) six or ten medals.[1] Further, some systems are forced rankings. If there are six wines in a flight, a judge must rank all six in order of

* The author thanks [].

[a] Contact author at [].

[1] Depending on the year, the CSF has awarded six (No Award, Bronze, Silver, Gold-, Gold, and Gold+) or ten (No Award, Bronze–, Bronze, Bronze+, Silver-, Silver, Silver+, Gold–, Gold, and Gold+) ordered medals. The author holds a WSET Level III certification.

relative preference. *Liquid Assets* and *San Francisco FOG* are examples of tasting groups who employ that approach. See reviews and comparisons of rating systems in Cicchetti & Cicchetti (2014), Kliparchuck (2013), and Veseth (2008).

Although judges and critics focus on the wine in the glass, much research shows that the ratings that tasters assign are affected by other factors. Some of the factors are supported by literature that is cited below. Other factors described below are reported as anecdotal and, when no literature is cited, those other factors are intended as hypotheses that remain to be tested.

Stochastic Ratings

First, although wine ratings are not merely random, evidence that ratings are stochastic is abundant in the wine-related academic literature and trade press. Bodington (2017, 2020) summarizes and cites four experiments with blind replicates, more than 20 other evaluations that find uncertainty in ratings, and two texts that explain the neurological, physiological, and psychological reasons for variance in the ratings that the same judge assigns to the same wine. The stochastic nature of wine ratings is not unique. Kahneman *et al.* (2021, p. 80-86, 215-258) describe variance in wine ratings and in many other areas of human judgement including physicians' diagnoses, forensic experts' fingerprint identifications, and judges' sentencings of criminals.

The wine-related literature cited above supports several findings. A rating observed is one draw from a latent distribution, it is one instance of in some cases many potential instances. Ratings are heteroscedastic, so the distribution of ratings on a wine is wine- and judge-specific and different judges' ratings on the same wine are not identically distributed (ID). Some judges assign ratings more consistently than others, and some wines are more difficult to rate consistently than others. Research attempts to predict ratings from physiochemical properties have struggled to obtain statistical significance. Experiments with blind replicates show that, on average, the standard deviation of the rating that the same judge assigns to the same wine within a flight is approximately 1.3 out of 10 potential rating categories. And while some judges independently assign ratings that correlate well with each other, about 10% of CSF judges assign ratings that are indistinguishable from random assignments.

Although most ratings are assigned by judges prior to any discussion of the subject wines, pre-rating discussion sometimes occurs among panelists and some competitions require an initial rating, then discussion, and then a post-discussion rating. According to Taber (2005, p. 300-301), discussion of the wines took place during the tasting at the 1976 Judgement of Paris. The CSF is an example of a competition in which judges assign an initial rating, discuss the wines with other judges, and then assign another post-discussion rating. Both sets of ratings are reported to CSF officials and the author is not aware of any correlation or other comparisons made by the CSF. Judges can influence each other so post-discussion ratings may not be statistically independent (I). When combined with the heteroscedasticity above, post-discussion ratings may therefore not obtain the statistical ideal of independent and identically distributed (IID) observations.

Anchoring, Expectation, Serial Position, and Commercial Biases

The score-based rating systems noted above also assign categories of quality or award to score thresholds and ranges. For example, the OIV system sets score thresholds for Bronze, Silver, and Gold medals at scores of 80, 85 and 90 respectively.[2] With a sample of 8,400 ratings, Bodington and Malfeito (2018) showed spikes in the frequencies of scores assigned just below those thresholds. Thus, some judges appear to anchor at OIV's category thresholds.

In addition to anchoring scores to category thresholds, there is anecdotal evidence of sequential anchoring. In a taste-and-score sequential protocol, a judge may assign a rating to the first wine and then rate the remaining wines "around" that anchor. A lag structure may also exist in which a judge rates around some composite of the most recent wines. The upper and lower bounds on ratings, whether numerical or categorical, may then merely bound a judge's assessments of relative preference.

[2] The complete OIV award system is Bronze to wines with a mean score of at least 80 points (up to a maximum of 25% of all prized wines including Gold and Silver), Silver to wines with a mean score more than 84 points (up to a maximum of 12% of all wines entered), and Gold to wines with mean scores over 90 points (up to a maximum of 6% of all wines entered). A fourth medal, Great Gold, is awarded by a Grand Jury to the best wine in each of several categories (up to a maximum of 25% of the number of Gold medals).

Much research shows that judges' expectations affect the ratings they assign. Ashton (2014) found that judges assigned higher ratings to wines from New Jersey when told the wines were from California and lower ratings to wines from California when told the wines were from New Jersey. On that evidence, regardless of actual quality, an expectation of good quality may lead to a central tendency in ratings within whatever range of scores or categories indicates good quality. In addition, information provided about wines may alter expectations and ratings. For example, the pre-printed forms provided to CSF judges list the grape variety, vintage, alcohol by volume and residual sugar of a wine next to spaces where the judge writes in a comment and then a rating.[3] Whether or not such judgments should be represented as "blind" is open to debate. Whether or not having that information affects judges' ratings remains to be tested.

Some critics and competitions employ a sequential, step-by-step or taste-and-rate protocol. A critic or judge tastes a wine and assigns a rating, then tastes the next wine and assigns a rating to that wine, and so on. The Judgement of Paris, the CSF, and many publishing critics employ a sequential protocol.[4] The possibility that ratings assigned during taste-and-rate protocols are affected by serial position, rather than the intrinsic qualities of the wines and judges, is difficult assess and rule out. Serial position bias may occur in wine competitions due to palate fatigue, rest breaks, meal breaks, physiological and psychological factors.[5] There are anecdotal reports from judges who say there is temptation to assign a high rating to a dry and high-acid wine because it is refreshing in a sequence just after several off-dry and alcoholic wines. U.C. Davis' class for potential wine judges warns of position bias affecting differences in ratings due to the sequence of

[3] Form provided to the author by the CSF on July 16, 2019.

[4] In contrast, some competitions employ an "open" protocol in which a flight is poured and judges can taste and re-taste the wines in any order and frequency. *Liquid Assets* and *San Francisco FOG* follow that open protocol.

[5] Serial position bias is common in many fields of judging. de Bruin (2005) examined singing and figure skating competition results and found position bias in both step-by-step and end-of-sequence sequential judging protocols.

wines, breaks and lunch.[6] Filipello (1955, 1956, 1957) and Filipello and Berg (1958) conducted various tests using sequential protocols and found evidence of primacy bias. Mantonakis *et al*. (2009, p. 1311) found that "high knowledge" wine tasters are more prone than "low knowledge" tasters to primacy and recency bias. The sequence of wines tasted at the 1976 Judgment of Paris has never been disclosed so what effect position bias may have had on the results remains unknown.[7]

Profit-seeking may cause, in some cases and when ratings are not assigned blind, a commercial bias in ratings. Gray (2014, 2022) reports that wine ratings have been leaked to wine sellers in advance of publication, retailers with non-public information have cornered markets for some wines, and that ratings have been influenced by vintners' payments for advertising, sponsorships of tables at events, and other "pay to play" actions. Teeter (2014) and Siegel (2019) assert that wine critics' acceptances of free samples, dinners, and trips affect both what they write about a wine and the rating that they assign to a wine. Gregutt (2022) asserts that some critics inflate scores to get publicity for themselves. Those practices decouple wine ratings from the intrinsic qualities of wine in a glass and they may also affect research concerning the relationship between wine ratings and prices. Those possibilities are mentioned here for completeness and the author is not aware of any published assessment of the extent of such practices.

III. Conclusion

Wine judges assess wines and assign ratings that are discrete and within bounded sets of scores, ordered categories, or ranks. But much evidence shows that, although they are not merely random, those assignments are both stochastic and heteroscedastic. Those assignments may also be affected by anchoring, expectations, serial position, and commercial biases.

[6] The author took the class and test for potential CSF judges at UC Davis.

[7] The Judgment's tasting protocol was sequential taste-and-score. The author confirmed, in email communications with both Mr. Taber and Mr. Spurrier, that the sequence of pour has never been disclosed.

References:

Ashton, R.H. (2014). Nothing good vver came from New Jersey: Expectations and the sensory perception of wine. *Journal of Wine Economics*, 9(3), 304-319.

Bodington, J. C. (2017). The distribution of ratings assigned to blind replicates. *Journal of Wine Economics*, 12(4), 363-369.

Bodington, J. C. (2020). The latent distribution of a rating observed. AAWE Working Paper No 259, October 2020, 15 pages.

Bodington, J., and Malfeito-Ferreira, M. (2017). The 2016 wines of Portugal challenge: General implications of more than 8400 wine-score observations. *Journal of Wine Research*, 2(4), 313-325.

Cicchetti, D. & Cicchetti, A. (2014). Categorizing a wine rating scale: 2, 3, 4, or more: Is there one we should go for? *Journal of Business and Economics*, 5(8), 1199-1204.

de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, 118, 245-260.

Filipello, F. (1955). Small panel taste testing of wine. *American Journal of Enology*, 6(4), 26-32.

Filipello, F. (1956). Factors in the analysis of mass panel wine-preference data. *Food Technology*, 10, 321-326.

Filipello, F. (1957). Organoleptic wine-quality evaluation II. Performance of judges. *Food Technology*, 11, 51-53.

Filipello, F. and H.W. Berg (1958). The Present Status of Consumer Tests on Wine. Presentation to the Ninth Annual Meeting of the American Society of Enologists, Asilomar, Pacific Grove, California, June 27-28, 1958.

Gray, W.B. (2014). Wine Spectator, Advocate can now legally sell a 90-point rating. The Gray Report, 8 September 2014, https://blog.wblakegray.com/2014/09/wine-spectator-advocate-can-now-legally.html, accessed 24 January 2022.

Gray, W.B. (2022). Lisa Perrotti-Brown makes some interesting accusations, and maybe solves the Wine Advocate sake mystery. The Gray Report, 21 January, 2022, https://blog.wblakegray.com/2022/01/lisa-perrotti-brown-makes-some.html, accessed 24 January 2022.

Gregutt, P. (2022). Don't look up! Inflated scores are attacking the wine industry. PaulG on Wine, 5 February 2022, https://www.paulgwine.com/lets-discuss/dont-look-up-inflated-scores-are-attacking-the-wine-industry, accessed 4 March 2022.

Kahneman, D., Sibony, O., and Sunstein, C.R. (2021). Noise, a flaw in human judgement. New York, Little, Brown Spark, Hachette Book Group, 454 pages.

Kliparchuck, K. (2013). What's wrong with wine ratings? MyWinePal, April 8, 2013.

Mantonakis, A., P. Rodero, I. Lesschaeve and R. Hastie (2009). Order in choice: Effects of serial position on preferences. *Psychological Science*, November 2009, 20(11), 1309-1312.

Siegel, J. (2019). Is ”pay to play” wrecking wine criticism?, Wine Curmudgeon, 25 April, 2019, https://www.winecurmudgeon.com/is-pay-to-play-wrecking-wine-criticism/,accessed 26 January 2022.

Taber, G.M., (2005). Judgement of Paris. New York, Scribner, 326 pgs.

Teeter, A. (2014). How a ratings system designed to help consumers is tearing the $40 billion dollar wine industry apart.” VinePair, 13 November 2014, https://vinepair.com/wine-blog/wine-ratings-industrial-complex/, accessed 26 January 2022.

Veseth, M. (2008). Wine by the numbers. The Wine Economist. February 9, 2008.