Journal of Wine Economics (2022), **17**, 345–351
doi:10.1017/jwe.2022.53

ARTICLE

Stochastic error and biases remain in blind wine ratings

Jeffrey Bodington

Bodington & Company, San Francisco, CA, USA
Email: jcb@bodingtonandcompany.com

Abstract

Analyses and aggregations of the ratings that wine critics and judges assign to wines are made difficult by stochastic error and biases that remain even when wines are assessed blind to price, label, capsule, and closure. Stochastic error is due to the partially random nature of ratings. Cognitive and omitted-variable biases are due to anchoring, expectation, serial position, commercial, and other factors. Differences in decanting, filtering, aeration, and temperature can also affect ratings.

Keywords: bias; judge; random; ratings; statistics; wine

JEL Classifications: A10; C00; C10; C12; D12

I. Introduction

The ratings that critics and judges 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 price, label, capsule, and closure), several sources of potential error and bias remain. Even when a wine is assessed blind, the rating assigned may be influenced by factors that are not in the glass.

Section II of this short article is a review of the findings that blind wine ratings are uncertain and subject to stochastic error. Section III is a review of findings that anchoring, expectation, serial position, commercial, and non-taste non-smell sensory (sight, sound, touch) factors may induce cognitive and omitted-variable biases in blind ratings. Section IV is a review of findings that differences in physical preparation (decanting, filtering, aeration, temperature) can also be omitted variables that affect ratings. Conclusions and implications follow in Section V.

II. Stochastic errors

The uncertainty surrounding the ratings that judges assign, even when assigned blind, is old news. Without specific reference to wine ratings, Saal, Downey, and Lahey (1980) reviewed the history of variability in judgment-related ratings dating from

1909. Focusing on wine ratings, Filipello (1955, 1956, 1957), Filipello and Berg (1958), Tish (2004), Hodgson (2008a, 2008b), Ashton (2012, 2013), and Bodington (2012, 2017) published results showing that there is variance in the rating that a judge assigns to a wine. Ough and Baker (1961), Goldberg (1991), Castriota, Curzi, and Delmastro (2012), Goode (2014), Shepherd (2018), Bodington (2020), and Glancy (2020) published results showing that variance in ratings can change from wine to wine (for the same judge) and from judge to judge (for the same wine).

None of the literature cited previously means that wine ratings are merely random, but it does show that such ratings are both uncertain and heteroscedastic. Those findings are not unique to wine. Kahneman, Sibony, and Sunstein (2021, pp. 80–86, 215–258) describe heteroscedasticity in other areas of judgment, including diagnoses by physicians, fingerprint identifications, and sentencings of criminals by judges.

III. Cognitive and omitted-variable biases

Although judges and critics focus on the wine in the glass, evidence shows that blind ratings are affected by factors that are not in the glass. Some of those factors are indicated by literature that is cited next. Other factors described next are reported as anecdotal and thus hypotheses that remain to be tested.

A. Anchoring

Many score-based rating systems assign categories of quality or award to score thresholds and ranges. De Long (2006) describes ten well-known, score-based wine rating systems that have score ranges for different categories of quality. For example, the International Organization of Vine and Wine (OIV, 2021) prescribes scoring between 0 and 100 along with thresholds for Bronze, Silver, and Gold medals at scores of 80, 85, and 90, respectively. For 8,400 ratings according to the OIV system, Bodington and Malfeito-Ferreira (2017) showed spikes in the frequencies of scores assigned just below those thresholds.

Even without quality categories, evidence shows that critics and judges favor and avoid certain scores. For a competition that prescribed scores between 50 and 100, Bodington (2017) showed spikes in frequency at nearly every 5-point interval. Chaudhary and Siegel (2016) reported a sharp increase in scores of 90 and higher published in anonymous "major wine magazines." Hunt (2013) found spikes in the frequencies of the scores assigned by Jancis Robinson, Robert Parker, and others. While some of the results reported by Chaudhary and Siegel (2016) and Hunt (2013) may be due to sample bias, the combination of research cited here shows that some judges appear to anchor scores about categorical or psychological thresholds. Scores that appear to be cardinal may be more accurately interpreted as ordinal.

B. Expectations

Much research shows that judges' expectations affect the ratings that they assign. Ashton (2014) showed 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.

Several aspects of expectations remain to be tested. The pre-printed forms provided to California State Fair (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.[1] Whether or not such judgments should be represented as "blind" is open to debate, and that information may affect ratings. Furthermore, even when blind to all information about a wine, a judge's expectation of overall good quality based on the assumption that vintners enter their good and not so good wines in competitions may lead to a central tendency in ratings within whatever range of scores or categories indicates good quality.

C. Sequential position

In contrast to flights in which wines can be reassessed, sequential or taste-then-rate protocols are common. The Judgment of Paris, the CSF, and many publishing critics employ sequential protocols.

Serial position bias may occur in sequential tastings due to carryover, palate fatigue, rest breaks, meal breaks, physiological, and psychological factors.[2] 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. UC Davis' class for potential wine judges warns of position bias due to the sequence of wines, breaks, and lunch.[3] 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" wine 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.[4]

Other forms of position bias are possible. There is anecdotal evidence that, 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. In addition to the effects of stochastic error discussed in Section II, that lag structure could cause a resulting set of scores to violate transitive axioms of equality and inequality.

D. Commercial

Accusations that money and favors affect critics' writings and ratings are common in the wine-trade press.

Even when wines are assessed blind, some assert that commercial considerations affect judges' ratings. Gregutt (2022) wrote that some critics inflate scores to get

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

[2]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.

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

[4]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.

publicity for themselves. Gray (2013) reports that some competitions encourage judges to assign high scores and medals, and judges have told this author that competition officials asked them to assign more gold medals to increase current-entrant satisfaction and future submissions. Although those reports indicate a potential for commercial bias in some cases, to date, a documented analysis of such bias does not appear to have been published.

E. Other senses: Sight, sound and touch

Wine assessment is more sensory than just smell and taste. Sight, sound, and touch may affect ratings too. Chaudhary and Siegel (2016) showed that red wines tend to get higher ratings than white wines. Seeing the color of wine alone may affect expectations and thus ratings. Spence, Velasco, and Knoeferle (2014) showed that the color of the light in the tasting room and the type of music played affected the ratings assigned by over 3,000 novice tasters. North (2012) and Wang and Spence (2017) showed that background music can affect novices' and wine professionals' wine descriptors, purchases, and ratings. Campo, Reinoso-Carvalho, and Rosato (2021) reviewed the literature concerning how tasting wine is an experience in which taste, smell, vision, sound, and touch interact. Regarding touch, the quality of glassware was shown to affect a taster's perceptions of wine quality.

IV. Physical preparation: Decanting, filtering, aeration, and temperature

The physical preparation of wine, not yet in a glass, can alter what is in the glass. While physical preparation may not affect the relative ratings assigned by judges on a panel tasting from the same bottle at about the same time, it may affect the ratings assigned by judges at different times and/or after different preparations.

Although it is obvious that decanting can remove sediment and critics including Rosenthal (2008) argue the merits of filtering, no trials appear to have yet examined their potential effects on ratings. Wollan, Pham, and Wilkinson (2016) showed that exposure due to active aeration, or merely time in an open glass, enables evaporation of ethanol and other volatiles, including hydrogen sulfide, that "significantly influence the perception of wine attributes." Wollan, Pham, and Wilkinson did not examine the effects of aeration on short-term oxidation. Master of Wine Canterbury (2014) and Fox (2016) conducted informal blind trials with aerators and found no differences between aerated wines and wines poured into glasses and left to stand for a few minutes. Much is also written in the trade press about the best serving temperatures for various wines. Campo, Reinoso-Carvalho, and Rosato (2021) cite literature showing that temperature does affect mouthfeel and tasters' perceptions of aromas.

V. Conclusion and implications

Even when wines are assessed blind (to price, label, capsule, and closure), published research shows that the ratings that critics and judges assign to wines may be influenced by noise and biases, as shown in Figure 1.

Functional forms that treat ratings as if they are deterministic, or uncertain but identically distributed, are misspecifications of the uncertain and heteroscedastic

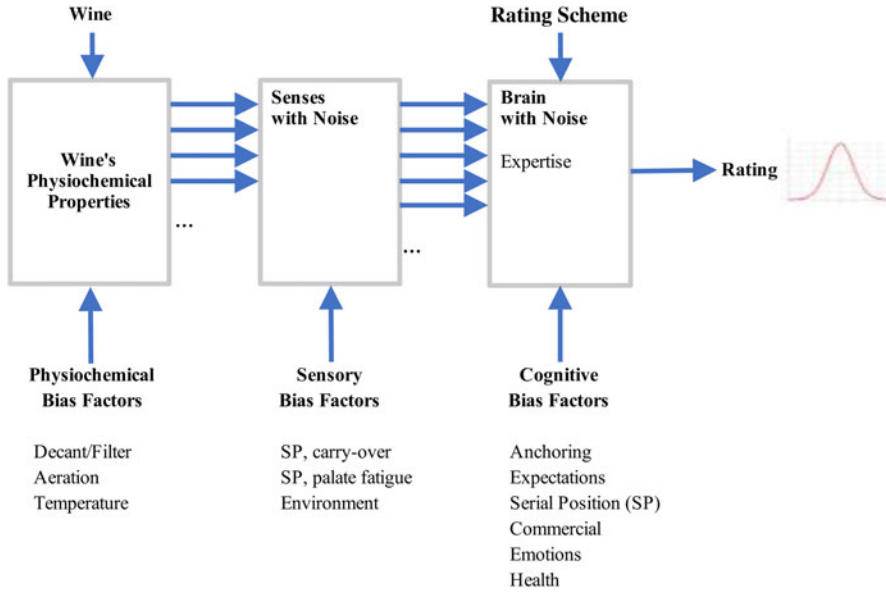

Figure 1. Summary of noise and biases that remain in blind ratings.

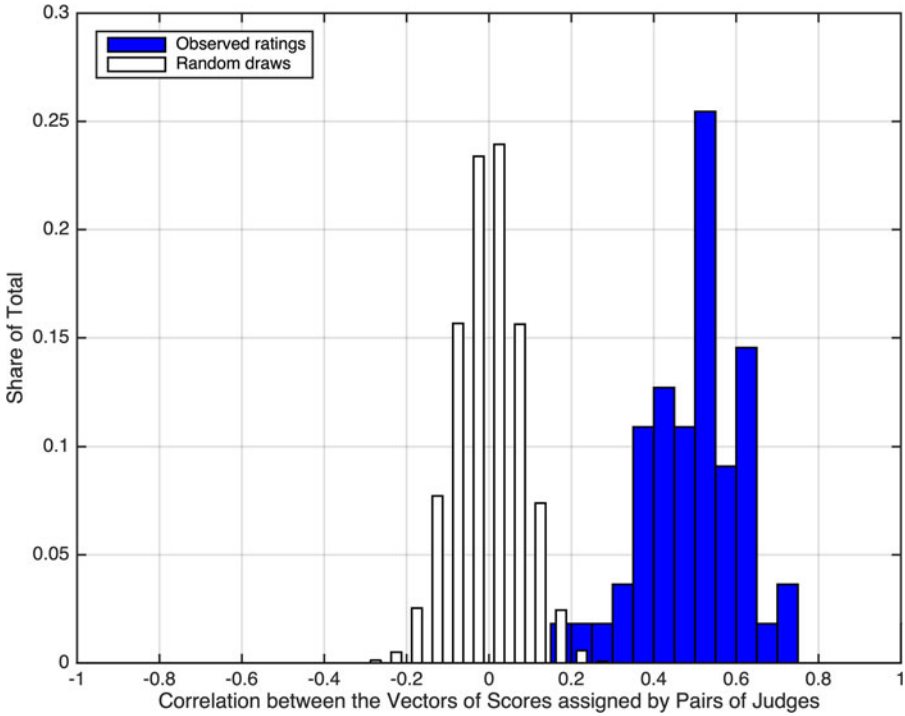

Figure 2. Frequency distribution of correlations between pairs of judges' ratings.

nature of ratings. The ratings that a judge assigns to a series of wines may not comply with the transitive axioms of equality and inequality. Analyses of scores as if they are cardinal may miss anchoring and cognitive biases that make them ordinal. Expectation, serial position, commercial, and sight-sound-touch sensory factors may induce cognitive and omitted-variable biases that cause a judge's rating to differ from a rating of only what is in the glass. In addition to stochastic error, ratings on blind replicates may be differentiated by carryover, palate fatigue, aeration, and temperature. Further, a central tendency in expectations and other factors may alter the null hypothesis that ought to be employed in tests of statistical significance.

While stochastic error and biases make an analysis of ratings difficult, ratings data are not merely random or impenetrable. For example, using 2019 CSF data from Bodington (2020), Figure 2 shows that the correlations between vectors of judges' ratings on the same wines concentrate between 0.3 and 0.7.

Acknowledgments. The author thanks an anonymous reviewer for insightful and constructive comments.

References

Ashton, R. H. (2012). Reliability and consensus of experienced wine judges: Expertise within and between? *Journal of Wine Economics*, 7(1), 70–87.

Ashton, R. H. (2013). Is there consensus among wine quality ratings of prominent critics? An empirical analysis of red Bordeaux, 2004–2010. *Journal of Wine Economics*, 8(2), 225–234.

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

Bodington, J. C. (2012). 804 Tastes: Evidence on randomness, preferences and value from blind tastings. *Journal of Wine Economics*, 7(2), 181–191.

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

Bodington, J. C. (2020). Rate the raters. *Journal of Wine Economics*, 15(4), 363–369.

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*, 28(4), 313–325.

Campo, R., Reinoso-Carvalho, F., and Rosato, P. (2021). Wine experiences: A review from a multisensory perspective. *Applied Sciences*, 11(10), 4488. doi.org/10.3390/app11104488

Canterbury, C. (2014). What does a wine aerator do? Kendal Jackson. Available at https://www.kj.com/blog/is-an-aerator-necessary-to-enjoy-wine (accessed May 22, 2022).

Castriota, S., Curzi, D., and Delmastro, M. (2012). Tasters' bias in wine guides' quality evaluations. American Association of Wine Economists, Working Paper No. 98, February. Available at https://wine-economics.org/wp-content/uploads/2012/10/AAWE_WP98.pdf.

Chaudhary, S., and Siegel, J. (2016). Expert scores and red wine bias: A visual exploration of a large dataset. *The Wine Curmudgeon*, October 24. Available at https://winecurmudgeon.com/expert-scores-red-wine-bias/.

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

De Long (2006). How we rate wines (and other things). *De Long Wines blog*, n.d. Download at http://www.delongwine.com/how_we_rate_wines.pdf (accessed April 22, 2022).

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.

Fox, S. (2016). Putting aerators to the test. *Minnesota Uncorked*, July 24. Available at https://www.minnesotauncorked.com/putting-aerators-to-the-test/ (accessed May 22, 2022).

Glancy, D. (2020). Do points matter? San Francisco Wine School, May 2020.

Goldberg, H. G. (1991). Pinning medals on California wines. *The New York Times*, December 11. Available at https://www.nytimes.com/1991/12/11/garden/pinning-medals-on-california-wines.html.
Goode, J. (2014). *The Science of Wine*, 2nd edition. London: Octopus Publishing Group, Ltd.
Gray, W. B. (2013) Wine competitions: For whom the medals toll? *Palate Press*, June 16. https://www.palatepress.com/wine-competition-for-whom-the-medals-toll/ (accessed March 26, 2022).
Gregutt, P. (2022). Don't look up! Inflated scores are attacking the wine industry. *PaulG On Wine*, February 5. Available at https://www.paulgwine.com/deep-dive/dont-look-up-inflated-scores-are-attacking-the-wine-industry (accessed March 4, 2022).
Hodgson, R. T. (2008a). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.
Hodgson, R. T. (2008b). On rating wines with unequal judges. *Journal of Wine Economics*, 3(2), 226–227.
Hunt, A. (2013) What's in a number? Part the second. *Jancis Robinson, The Purple Pages*, August 5. Available at https://www.jancisrobinson.com/articles/whats-in-a-number-part-the-second.
International Organisation of Vine and Wine (2021). OIV standard for international wine and spiritous beverages of viticultural origin competitions, edition 2021. Download at https://www.oiv.int/public/medias/7895/oiv-patronage-competition-norme-ed-2021.pdf (accessed May 19, 2022).
Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgement*. New York: Little, Brown Spark, Hachette Book Group.
Mantonakis, A., Rodero, P., Lesschaeve, I., and Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science*, 20(11), 1309–1312.
North, A. C. (2012). The effect of background music on the taste of wine. *British Journal of Psychology*, 103(3), 293–301.
Ough, C. S., and Baker, G. A. (1961). Small panel sensory evaluations of wines by scoring. *Hilgardia*, 30(19), 587–619.
Rosenthal, N. (2008). *Reflections of a Wine Merchant*. New York: North Point Press.
Saal, F. E., Downey, R. G., and Lahey, M. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413–428.
Shepherd, G. (2018). *Neuroenology: How the Brain Creates the Taste of Wine*. New York: Columbia University Press.
Spence, C., Velasco, C., and Knoeferle, K. (2014). A large sample study on the influence of the multisensory environment on the wine drinking experience. *Flavour*, 3(8), https://doi.org/10.1186/2044-7248-3-8.
Tish, W. R. (2004). Industry forum: Wine ratings 2 of 3: Ten reasons we all lose when numbers dominate the marketplace. *Wine Business Monthly*, Dec. Available at https://www.winebusiness.com/wbm/?go=getArticleSignIn&dataId=36265.
Wang, J., and Spence, S. (2017). Assessing the influence of music on wine perception among wine professionals. *Food Science & Nutrition*, 6(2), 295–301.
Wollan, D., Pham, D-T., and Wilkinson, K. (2016). Changes in wine ethanol content due to evaporation from wine glasses and implications for sensory analysis. *Journal of Agriculture and Food Chemistry*, 64(40), 7569–7575.

Cite this article: Bodington, J. (2022). Stochastic error and biases remain in blind wine ratings. *Journal of Wine Economics*, 17(4), 345–351. https://doi.org/10.1017/jwe.2022.53