



ARTICLE

Crowdsourcing the assessment of wine quality: Vivino ratings, professional critics, and the weather

Orestis Kopsacheilis¹ , Pantelis P. Analytis², Karthikeya Kaushik³, Stefan M. Herzog⁴ , Bahador Bahrami⁵ and Ophelia Deroy^{5,6}

¹Technical University of Munich (TUMCS, SOM), Munich, Germany; ²University of Southern Denmark, Odense, Denmark; ³University of California, Berkeley, CA, USA; ⁴Max Planck Institute for Human Development, Berlin, Germany; ⁵Ludwig Maximilian University, Munich, Germany and ⁶University of London, London, UK

Corresponding author: Orestis Kopsacheilis; Email: orestis.kopsacheilis@tum.de

Abstract

Crowdsourcing platforms—such as Vivino—that aggregate the opinions of large numbers of amateur wine reviewers represent a new source of information on the wine market. We assess the validity of aggregated Vivino ratings based on two criteria: correlation with professional critics' ratings and sensitivity to weather conditions affecting the quality of grapes. We construct a large, novel dataset consisting of Vivino ratings for a portfolio of red wines from Bordeaux, review scores from professional critics, and weather data from a local weather station. Vivino ratings correlate substantially with those of professional critics, but these correlations are smaller than those among professional critics. This difference can be partly attributed to differences in scope: Whereas amateurs focus on immediate pleasure, professionals gauge the wine's potential once it has matured. Moreover, both crowdsourced and professional ratings respond to weather conditions in line with what viticulture literature has identified as ideal, but also hint to detrimental effects of global warming on wine quality. In sum, our results demonstrate that crowdsourced ratings are a valid source of information and can generate valuable insights for both consumers and producers.

Keywords: crowdsourcing; wine quality; wine aging; global warming; Vivino

1. Introduction

Consider a prospective wine buyer staring down the wine aisle of a supermarket, faced with a seemingly endless variety of options. What information can they use to pick a wine that they will enjoy when tasting is not an option?

Numerous factors go into making a good wine, from the terroir in which it was produced to the vinification methods employed by the winemaker. Producers in certain

wine regions, such as Bordeaux, go to great lengths to ensure a recognized standard of quality. Even then, weather conditions cause severe yearly fluctuations to the health and quality of the grapes that were used to make the wine and that are, therefore, important determinants of wine quality. Keeping track and integrating all this information is a daunting task for the casual wine consumer.

Traditionally, the opinions of influential wine critics such as Robert Parker and Jancis Robinson (JR) have provided a key source of information for prospective wine buyers. But relying on these reviews is not always easy, as the few well-known critics tend to be selective about the wines they taste. Further, many critics have monetized their reviews by running subscription-only websites. Thanks to the advent of the Internet, prospective wine buyers can now tap into a new source of social information to help them navigate this inference problem: crowdsourced ratings from large communities of wine consumers on platforms such as Vivino and Cellartracker.

Our goal in this paper is to assess the validity of crowdsourced ratings in the domain of wine. To this end, we first examine the correlations between the averaged crowdsourced ratings of amateur Vivino users and the ratings of several professional critics. Second, we evaluate how these ratings reflect weather fluctuations over recent years.

In matters of taste, the quality of a judge's opinion is often proxied by its similarity to that of other judges (e.g., Ashton, 2012). However, relatively little is known about how crowdsourced ratings compare with the ratings of professional wine critics. We construct and analyze a novel and rich dataset consisting of Vivino ratings for a portfolio of red wines from Bordeaux. We then match our dataset with the ratings of eight professional critics and perform correlation analysis, treating Vivino as an independent critic.

Recognizing the limitations of using consensus as the only metric for the quality of information, we complement our analysis by assessing whether crowdsourced ratings mirror a set of objective markers of wine quality. Namely, we investigate the relation between Vivino ratings and the weather conditions that were present during the year that grapes were grown and harvested (i.e., the wine's "vintage"). We do this by collecting climatic information from a local weather station and exploring whether the Vivino ratings are responsive to variation in weather conditions known to affect wine production.

Our correlation analysis suggests substantial consensus between averaged Vivino ratings and professional critics' judgments. Moreover, regressing averaged ratings on local weather conditions shows that both amateur and professional ratings respond to the impact of meteorological conditions in similar ways and in line with findings from viticulture research. We conclude by identifying two promising research directions and take first steps toward addressing them.

First, we find that despite the considerable agreement between crowdsourced and professional ratings, there are also systematic discrepancies. Our exploratory analysis suggests that these are partly due to differences in scope: Amateurs' ratings emphasize the immediate pleasure of drinking a wine, whereas professional critics focus more on the potential of a wine once it has matured.

Second, we demonstrate that crowdsourced ratings can yield important insights regarding the impact of climate change on wine quality and consumption. Our analysis shows that prolonged high temperatures have a detrimental effect on the subjective

quality ratings of both amateurs and professionals. This result suggests that the hitherto positive relationship in the northern hemisphere between higher temperatures and wine quality may already have been disrupted.

Overall, our analysis suggests that crowdsourced ratings are a valid source of information, yielding useful insights for consumers and producers alike.

II. Background and motivation

A. Crowdsourced ratings

Wine is a prime example of an “experience good”—its quality is learned only after consumption (Nelson, 1970). In principle, consumers can use various observable cues about wines—including price, label design, and awards won in international competitions—to overcome this deficit and infer quality (Drichoutis et al., 2017). However, these heuristic strategies are not always reliable. For example, in a meta-analytic study, Oczkowski and Doucouliagos (2015) found only a modest correlation between prices and subjective reports of quality (the weighted average of all estimates was 0.30), casting doubt on the dictum that “you get what you pay for”—at least for wine. Moreover, Hodgson (2008) examined judge reliability at a major U.S. wine competition and found that only about 10% of judges were able to consistently replicate their score within a single medal group. Thus, medals and prizes seem unreliable as a source of information.

In recent years, the Internet has offered prospective buyers a new source of social information that can be leveraged to inform their choice: crowdsourced online ratings (e.g., Chevalier and Mayzlin, 2006). Relying on the opinion of a large, relatively inexperienced crowd has shown promising results in domains such as economic forecasting (Jame et al., 2016), funding of entrepreneurial endeavors (Mollick, 2014), and medical diagnostics (Kurvers et al., 2023), to name just a few. To a large extent, the success of crowdsourcing can be attributed to the “wisdom of the crowds.” According to this principle, the judgment errors of different individuals tend to cancel each other out when their judgments are aggregated, resulting in an average error that tends to be smaller than that of a randomly chosen individual (see Surowiecki, 2005, for a popular book summarizing the benefits of this principle; Analytis et al., 2018; Müller-Trede et al., 2018, for applications in matters of taste).

In the world of wine, freely available crowdsourcing apps such as Vivino, CellarTracker, and Wine-Searcher have extended the task of wine evaluation to a large and heterogeneous network of amateur wine enthusiasts, potentially creating the conditions for crowd wisdom to be accrued. However, the quality of information in aggregated ratings can be corroded by social influence (e.g., Le Mens et al., 2018; Muchnik et al., 2013) or strategic manipulation (Luca and Zervas, 2016). Assessing the quality and properties of crowdsourced online ratings remains an open scientific question in numerous consumer domains, including wine.

In this study, we create a novel and rich dataset consisting of individual wine reviews from Vivino and analyze how they relate to those ratings from professional critics as well as how they respond to a set of weather variables. Founded in 2010, Vivino is—according to its webpage—the world’s most downloaded wine app, featuring millions of reviews of wines from around the world. We focus on a portfolio of red wines from

Bordeaux and track their Vivino ratings over time. Each wine is observed over 13 vintages, from the 2004 to the 2016 vintage. Critics' scores are obtained from en primeur events, at which critics and merchants are invited to taste wines from the barrel when they are just 6–8 months old.

B. Consensus and expertise

Ideally, the validity of judgments would be assessed on the basis of a set of objective criteria. In matters of subjective taste, however, such objectivity is hard to come by and researchers usually rely on alternative benchmarks. Consensus—typically measured by the degree to which judgments from experts correlate with each other—is arguably the most common such benchmark (Cicchetti, 2004, Ashton, 2012, 2013).¹

Even though consensus between experts has received considerable attention by the literature, relatively little is known regarding the consensus between judgments from expert critics and crowdsourced amateur ones. To our knowledge, there are three previous studies that focus on this relation. Oczkowski and Pawsey (2019) and Bazen et al. (2023) compared the impact of crowdsourced versus professional ratings on wine prices, while Gokcekus et al. (2015) focused on their relative influence on consumers. All three investigations seem to converge toward the conclusion that crowdsourced data are becoming increasingly influential. In this study, we make a more direct comparison between crowdsourced and professional ratings and shed light on the agreement between the two. Our dataset includes a sizeable overlap of wines reviewed by both Vivino amateurs and professional critics, making it particularly suitable for this type of analysis.

Despite its usefulness and ease of application, using consensus as the sole arbiter for evaluating the validity of a judgment has limitations. For example, in certain occasions the majority opinion has been shown to be systematically wrong (Galesic et al., 2018; Prelec et al., 2017). It has also been argued that disagreement can be a catalyst for enhancing knowledge. In the domain of peer-reviewed publications, for instance, editors sometimes select reviewers for their complementary perspectives (Weiss and Shanteau, 2004). In that respect, there is often a trade-off between validity (as proxied by consensus) and diversity of information (see also Broomell and Budescu, 2009).

Therefore, we complement our analysis with an alternative strategy for evaluating the validity of these crowdsourced judgments, taking advantage of a latent relationship between the quality of a wine and the weather conditions during the year of the harvest (i.e., the wine's "vintage"). Although the role of subjectivity in matters of taste cannot be overemphasized, some of the physical processes determining the quality of grapes can be objectively observed and measured. We thus assess the extent to which averaged Vivino ratings are sensitive to aspects of weather variability known to affect grape quality (i.e., temperature and rainfall at different points in the season). We further compare their responsiveness with that of professional critics.

¹The terms "decision similarity," "agreement," "inter-judge correlation," "reliability," and "concordance" are often used interchangeably with "consensus" in this literature.

The idea that judgments about quality contain both objective and subjective components is not new. Cicchetti (1991) drew attention to this duality in assessing the reliability of peer reviews, pointing out that the attributes for evaluating manuscripts “can be derived from either objective judgments (e.g., experimental design) or subjective ones (e.g., importance).” In the domain of wine, the notion that subjective ratings are partly governed by objective markers is summarized by Cardebat et al. (2014), who assumed that, besides subjective tastes, wine judgments have an objective component that is driven by the fundamentals of wine production, such as the quality of the soil, the producers’ skills, and—crucially to our analysis—weather conditions.

C. Weather and wine quality

Whether a wine’s quality can be assessed on the sole basis of objectively observable parameters such as the weather conditions has been a key question in the literature for the past 40 years. Ashenfelter’s seminal work in the 1980s and 1990s provided a highly successful econometric model for assessing the quality of Bordeaux vintages and predicting their prices in auctions based on the wine’s age and the weather conditions during the growing season (see Ashenfelter, 2008b, for an updated version). Often referred to as the “Bordeaux equation,” this model regresses a vintage-level price index (obtained from auctions of a specific wine portfolio) onto a set of weather variables and the wine’s age. The model has proven surprisingly effective at assessing the quality of Bordeaux vintages and predicting the prices of mature wines (Storchmann, 2012).

Inspired by Ashenfelter’s Bordeaux equation, we evaluate the responsiveness of averaged Vivino ratings to the same weather variables, namely, average temperatures and total rainfall during the preseason and the growing season, as measured by the local weather station at Merignac.

The wine region of Bordeaux is located in southwestern France, between 44.5° and 45.5°N. In such northerly latitudes, warmer growing seasons are expected to lead to higher fruit quality, which translates into better quality wine. Field evidence has thus far indeed confirmed that higher temperatures are beneficial for wine quality—often proxied by wine prices or winery revenue—in the relatively cooler climes of the northern hemisphere (Ashenfelter and Storchmann, 2010b; Jones et al., 2005). Even though global warming is likely to eventually harm the quality of grapes, there is no evidence for the detriments of excessive heat in wine regions of the northern hemisphere, with (Ashenfelter and Storchmann, 2010a) making a call for additional research on that issue.

With respect to precipitation, there is a consensus that rain during the last stage of the growing season, most notably in August, is detrimental for the health of grapes. Humidity during this sensitive period for berries can raise mildew pressure, which can cause rot from the inside out on thin-skinned tight clustered varieties (Matthews et al., 1987; Poni et al., 1993). There is less consensus regarding the effect of rain during the preseason (October–March), with some studies reporting on a positive effect (Ashenfelter, 2008b), while others find it to be not significant or even negative (Ashenfelter and Storchmann, 2010b).

III. Methods and results

For the amateur ratings, we collect our data from Vivino's public data for a portfolio of red wines from Bordeaux. As our focus was to examine the relationship between amateur and professional tastes, we compiled this portfolio based on the wines that feature in "Bordoverview" (<https://www.bordoverview.com/>)—a website reporting on the ratings from various professional critics provided at en primeur events for all Grand Crus and several "second wines" of Bordeaux. We restrict the dataset to those wine-labels for which we can find ratings for every year between 2004 and 2016.² Our initial dataset consists of all Vivino ratings for these wines that were available at the time of our data collection (July 2020). This amounts to 79,648 ratings for a total of 780 wines: 60 Chateaux observed over 13 consecutive years.³

Next, we match this dataset with the ratings from professional critics that were available in Bordoverview. Specifically, we focus on the ratings provided by the following six individual critics: James Suckling, Jancis Robinson, Jeff Leve, Neal Martin, Rene Gabriel, Tim Atkin, as well as Decanter and the Wine Advocate, two outlets summarizing the ratings of small groups of individual critics. We keep only wines that were reviewed by at least three of the aforementioned professional critics. The resulting matched dataset consists of 39,035 ratings for 371 wines from 41 chateaux. Older and younger vintages are equally represented through this matching process. Specifically, the number of observations per vintage from 2004 to 2016 is: [20, 23, 31, 26, 26, 30, 34, 34, 35, 29, 16, 32, 35], respectively. The maximum overlap is between Vivino and Decanter (360 matches); the median overlap between Vivino and a critic is 200 wine ratings. [Figure A1](#), in the Appendix, provides a summary of the number of wines per critic that we were able to match to a Vivino average.

A. Consensus analysis

We begin with a macroscopic view of the consensus between Vivino amateurs and those of Jeff Leve, an established professional critic specializing in the Bordeaux wine region. [Figure 1](#) compares Vivino ratings—averaged at the vintage level—with Jeff Leve's ratings. We collected Jeff Leve's vintage-level ratings from his website⁴. His vintage assessments are not based on the averages of individual wines; instead, they represent his general assessment of the wine quality for a specific year.

We found substantial resemblance between the two. For example, both sources agreed that the 2013 vintage was the worst in recent years and that the 2005 vintage was the best. Both of these claims are widely shared within the wine community.

²Although Vivino has only been in operation since 2013, the site includes reviews of vintages going further back in time. We chose the 2004 vintage as our starting point because this is the oldest vintage included in Bordoverview. Moreover, since vintages are often released in the market 2–3 years after being bottled, we anticipated a relative scarcity of reviews for vintages newer than 2016 to be available at the time of our data collection (July 2020).

³In our dataset, each chateau is represented by a single wine label and therefore the two terms are used interchangeably. A "wine" is defined at the level of a chateau (wine label) and a vintage, so that—for example—"Chateau d'Agassac Haut-Medoc 2004" and "Chateau d'Agassac Haut-Medoc 2005" represent two different wines in our dataset.

⁴<http://www.thewinecellarinsider.com>

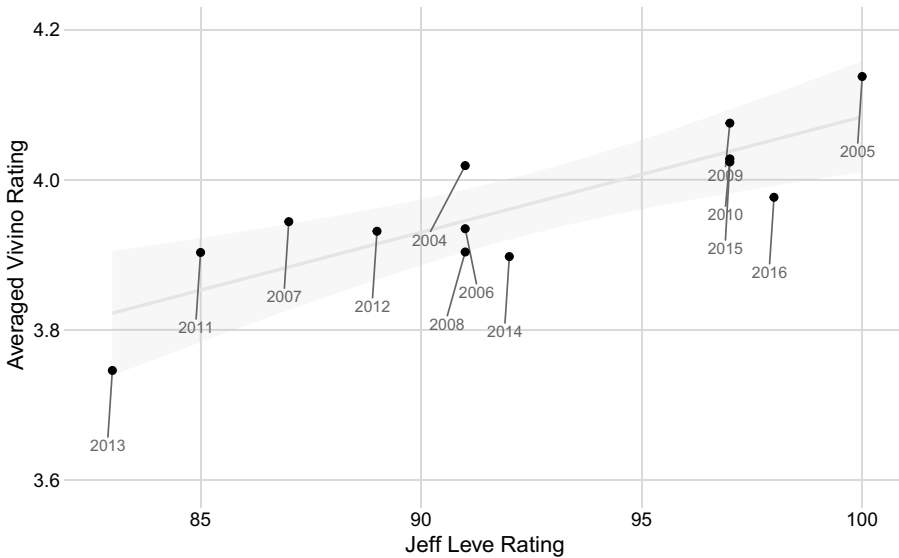


Figure 1. Vintage-level correlation between Vivino and Jeff Leve ratings.

Notes: Averaged Vivino ratings—at the level of the vintage—against Jeff Leve’s assessment of each vintage’s overall quality. For this figure we use the entire dataset of Vivino ratings and the raw ratings (i.e., no Z-transformations). Vivino ratings range from 1 to 5, with 1 being the lowest and 5 the highest possible score. Jeff Leve’s ratings range from 0 (lowest) to 100 (highest). The regression line’s coefficient is 0.014 (SE: 0.002) and is derived from the corresponding linear regression of Vivino over Jeff Leve’s vintage-level ratings and a constant. The line shows the ordinary least squares slope and its 95% confidence band.

For instance, St’éphane Derenoncourt, a French vigneron working as a consultant for numerous estates in Bordeaux, described producing the 2013 vintage as a “war against nature.” In contrast, the 2005 vintage has been described by some wine journalists as “majestic” (Asimov, 2021).

Next, we focus on the relationship between the tastes of Vivino amateurs and professional critics at the level of individual wines. Here, we treat averaged Vivino ratings as an independent critic.⁵ Figure 2 reports the two-way Pearson’s correlations (r) across pairs of critics (left) as well as the average correlation (\bar{r}) calculated by taking the arithmetic mean (right).

Two things are apparent from Figure 2. First, Vivino average ratings correlated substantially with the ratings of most professional critics, with a total correlation average of 0.40. Some critics, such as the Wine Advocate ($r = 0.50$) and Jeff Leve ($r = 0.48$), seem to be more in tune with the wine-loving crowd reviewing at Vivino than others, such as Decanter ($r = 0.16$), Jancis Robinson ($r = 0.36$), or James Suckling ($r = 0.37$).

Second, professional critics’ ratings still correlate more strongly with each other than with Vivino. Jeff Leve exhibits the overall highest average correlation ($\bar{r} = 0.63$), followed by Neal Martin ($\bar{r} = 0.62$), while Tim Atkin and Decanter have the lowest ($\bar{r} = 0.46$ and 0.49 , respectively).

⁵Unless specified otherwise, “average” always refers to the arithmetic mean.

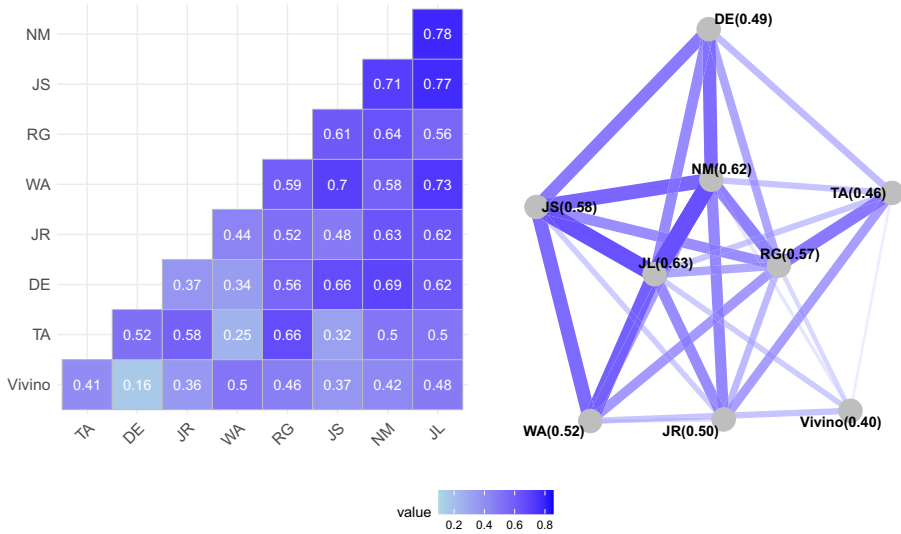


Figure 2. Correlations between wine raters.

Notes: Left: Visualization of correlation matrix with raters ordered in ascending order according to their average inter-correlation. Reported values correspond to Pearson’s correlation coefficients (r). Right: Correlation network with each rater represented as a separate node. The thickness of the edges is proportional to the strength of correlation between the judgments of two raters. Only edges corresponding to correlations of at least $r = 0.40$ are plotted. Color gradient is proportional to strength of correlation in both panels. Vivino: averaged ratings from Vivino users. DE: Decanter; JS: James Suckling; JR: Jancis Robinson; JL: Jeff Leve; NM: Neal Martin; RG: Rene Gabriel; TA: Tim Atkin; WA: the Wine Advocate.

We return to these points in Section IV A, where we take a closer look at these systematic differences between amateurs’ and professionals’ ratings.

B. Responsiveness to weather conditions

Here, we examine how averaged ratings from Vivino amateurs and professional critics reflect the weather conditions during a certain vintage. Inspired by Ashenfelter’s “Bordeaux equation” (Ashenfelter, 2008b), we study the responsiveness of these ratings on average temperature over the growing season (April–August), the average temperature in September, total rainfall in the preseason (October–March), and total rainfall in August. We also include a time trend, an index variable tracking the vintages in our dataset (from 2004 to 2016), to see if there is a linear tendency of average ratings to grow or diminish over the years. Table A1 in the Appendix provides key summary statistics on the variables used in our ordinary least squares regression analysis, while Figure A2 plots the distribution of ratings (for Vivino and professional critics) against each weather variable.

For our regression analysis, we treat our dataset as panel data, where each chateau (cross-sectional dimension) is observed over subsequent vintages (temporal dimension). The dependent variable of the Vivino model is constructed by averaging at the

level of the wine (i.e., a chateau in a given vintage) and then transforming these averages into Z-scores by subtracting the mean and dividing by the standard deviation. For the professional critics model, the process is the same but we add an additional step. Namely, we start by calculating Z-scores for each critic's ratings before averaging those Z-scores over the dataset. The reason for this additional step is that each critic uses their own rating system and it would be impossible to aggregate otherwise.

We implement a weighted least square multiple regression approach to examine how weather conditions affect perceived quality. Weights are proportional to the number of individual ratings from which each averaged rating was derived. The median number of ratings per averaged rating is 73 (IQR = 35, 136) for amateurs and 5 (IQR = 5, 6) for professionals. To account for the fact that the baseline quality of the wine can be different from one chateau to another, we use separate fixed effects for each chateau. However, as argued by Ashenfelter and Storchmann (2010b), we believe that—given the similarity of the wines planted in this region—the weather conditions can be expected to have similar effects across the wineries. We used Driscoll–Kraay standard errors, which are robust to both cross-sectional (cross-chateaux) and temporal (cross-vintages) dependence. The results of our regression analysis are displayed in Table 1.⁶

The signs of the coefficients of the weather variables tell a consistent story across both models. Higher average temperatures during the growing season have a beneficial impact on the subjective rating of wine quality for both amateurs and professionals.⁷ However, the significant negative coefficient of the average temperature in September suggests that the effect can be detrimental when high temperatures extend deep into the growing season. We return to this point in Section IV B.

Amateurs and professionals also agreed on the impact of rain. Averaged ratings reacted positively to rain preceding growth but negatively to rain in August, when the grapes mature.

Although amateurs and experts are overall very much in agreement, there are two noticeable differences that warrant attention. First, the intensity of the responsiveness to weather conditions captured by the size of the coefficients differs, with experts' ratings being more responsive. This might be in part due to differences in the variability of tastes within the two populations. The amateur crowd consists of thousands of raters, some of whom might have antithetical tastes, whereas experts' tastes are more likely to

⁶An implication of matching our initial dataset with that containing professional critics' ratings is that the resulting panel is unbalanced, as we do not always have at least three critics reviewing all vintages for each wine label. Table A2 in the Appendix reports on a similar regression model as that in Table 1 which focuses on ratings from Vivino ratings only. This allows us to compare the coefficients of the matched (unbalanced panel) with those from the unmatched (balanced panel) dataset. This analysis also allows us to consider a variant of the set of independent variables that is closer to the "Bordeaux equation" as it was first introduced by Ashenfelter. Reassuringly, the main findings of this section regarding Vivino ratings' responsiveness to weather also hold under this alternative analysis.

⁷Amateurs and professionals are in similar agreement with respect to the negative sign of the quadratic term of those average temperatures. Although this suggests that the relation between temperature and ratings may not be monotonic and higher temperatures would eventually backfire, the coefficients are not statistically significant. This is likely due to the fact that the observation period ($T = 13$) is too short to capture the inverse U-shape of this relation accurately.

Table 1. Determinants of wine ratings: Regressing ratings onto weather variables

	Model 1 Vivino	Model 2 Critics
Average temperature in growing season (April–August)	0.6268*** (0.1205)	0.8274*** (0.0917)
Average temperature in growing season squared	-0.4707 (0.2272)	-0.6010 (0.2976)
Average temperature in September	-0.1311** (0.0291)	-0.2139*** (0.0212)
Rainfall in preseason (October–March)	0.0006 (0.0007)	0.0028*** (0.0005)
Rainfall in August	-0.0055** (0.0013)	-0.0043** (0.0014)
Time trend	-0.0313* (0.0120)	0.1177*** (0.0077)
Observations	371	371
Number of groups	41	41
F-statistic	40.11	86.24
R ² for within model	0.3989	0.6291

Notes: The averaged, standard-normalized rating of a wine (Z-score) is regressed onto climate variables which have been centered to their mean. Model 1: Ratings from Vivino amateurs. Model 2: Ratings from professional critics. The regression models include fixed effects at the level of the chateau and weights proportional to the number of ratings included in the calculation of each average. Each wine in our data set is uniquely identified by a chateau and a vintage. Average temperatures are measured in °C. Rainfall is measured in mm of water accumulated over the entire period. Driscoll-Kraay (vintage- and chateau-clustered) standard errors are in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

be aligned. This asymmetric variation would also explain the discrepancy between the fit of these two models as captured by the R^2 statistic (R^2 -within = 0.411 for Vivino amateurs; R^2 -within = 0.6109 for professional critics). Second, the sign of the time trend suggests that amateurs and professionals have—on average—different attitudes toward younger vintages. Amateur tasters tend to rate more favorably older vintages, whereas the opposite is the case for professional critics. This point is further discussed in Section IV A, where we explore the underpinnings of this apparent difference in tastes between the two groups.

IV. Exploratory analysis and discussion

A. Differences between Vivino amateurs and professional critics

Our consensus analysis in Section III A revealed that Vivino ratings correlated substantially with those of professional critics. Besides the vintage-level agreement with Jeff Leve, the average correlation with professional critics was well within the range of those reported between critics in other studies (Ashton, 2012; Stuen et al., 2015) and even comparable with the consensus among experts in other domains, such as in clinical psychology (Ashton, 2012). However, it was lower than the average correlation among critics in our dataset as well as in other studies using en primeur data (Masset

et al., 2015). What can account for the systematic discrepancy between amateurs and professionals?

To address this question, we followed a “lead” from the analysis in Section III B—which revealed antithetical views between amateurs and professionals with respect to attitudes toward younger vintages. One possible interpretation of this asymmetry is that the two crowds differ in the scope of their evaluation, with amateurs focusing on the immediate pleasure of consuming the product, but critics judging how wines will develop over time.

Many wines have aging potential, reflecting ongoing chemical processes that persist well after fermentation ends (Goode, 2005), and would thus have improved had the consumer not stopped the maturation process by opening the bottle. This aging potential is particularly pronounced for red wines from Bordeaux, which are typically rich in tannins and may taste astringent and unpleasant if drunk at a young age.

This hypothesis can explain the negative time trend observed for amateurs (but not professionals) in the analysis in Table 1. More recent vintages have had less time to mature (as we only collected Vivino ratings up to 2016) and are therefore, keeping everything else constant, judged more harshly if assessed primarily based on their immediate quality.

Figure 3 makes this last point clearer. It plots averaged Vivino ratings against the age of the vintage at the time it was consumed. The variable “Age” is calculated as the difference in years between the harvesting of the grapes and the posting of the review ($M = 7.26$, $Med = 7$, $Q1 = 5$, $Q3 = 9$). All 13 panels—each tracking this effect for a different vintage—show clear evidence for an upward slope, suggesting that reviews posted later in the wine’s maturation cycle tend to be more generous.

Table A2, in the Appendix, further illustrates this point. There, we repeat a version of the analysis reported in Table 1, but add the variable “Average Age” to the set of regressors. Verifying the visual impression of Figure 3, we find the age-coefficient to be significantly positive, suggesting that Vivino ratings are not fully accounting for the wine’s potential. In line with this interpretation is also the fact that the coefficient of the time trend is no longer negative after controlling for the age of the wine when the bottle was opened. This implies that Vivino users do not find the quality of younger vintages inferior, they just did not fully account for their potential when rating them prematurely.

Our dataset does not allow us to test the same effect for experts—we only have access to en primeur reviews given before the wines are bottled. However, we can refer to the Global Wine Score team (The Global Wine Score, 2017), which has conducted a similar analysis for experts, using re-notations that capture critics’ adjustments to their en primeur assessment in subsequent years. If critics did not account for the wine’s improvement with age, we would expect to see, on average, systematic positive adjustments. Instead, the report finds that the net adjustments, averaged across all wines and across all critics for each subsequent year, are either zero or slightly negative (though not statistically significantly so). This implies that critics on average account for a wine’s maturation potential in their en primeur assessments.

In the past, differences in the subjective evaluation of experience goods between amateurs and expert critics have been typically attributed to differences in taste (Holbrook, 1999). In line with that perspective, some contributions have emphasized

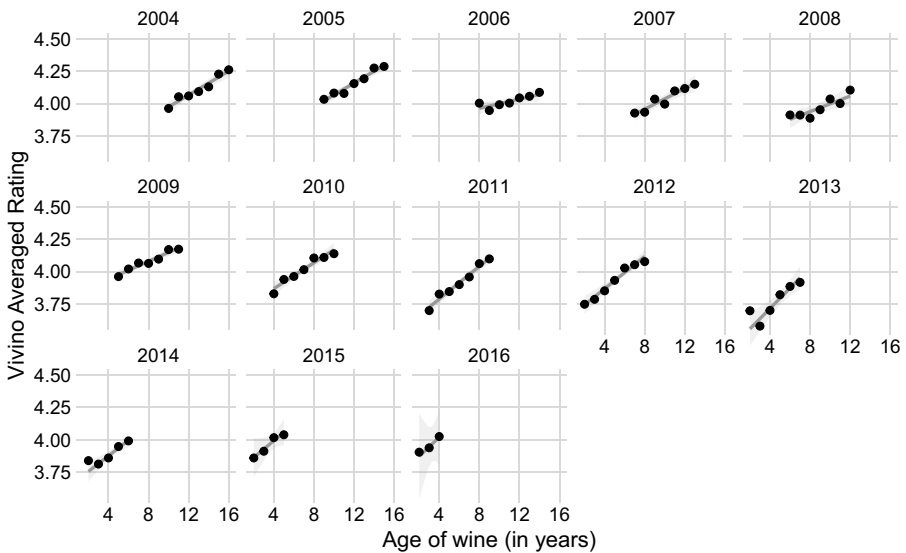


Figure 3. Vivino ratings by age of the vintage at the time consumed.

Notes: Vivino ratings, averaged at the level of the age of each wine when it was consumed. We proxy “age” by taking the difference (in years) between the date of the review and the wine’s vintage (year in which the grapes were harvested). Lines show weighted ordinary least squares slopes and their 95% confidence band.

the role of different levels of experience in taste divergence (see Goldstein et al., 2008 for wine; McAuley and Leskovec, 2013 for beer). Our analysis suggests a new mechanism that can also account for part of this discrepancy—at least for red wine—namely, that the two crowds seem to differ in the scope of their evaluation.

Future research aiming to disentangle the underpinnings of this difference in tastes can yield important insights for the understanding of preferences, with pertinent applications to recommender systems. This research would need to control for additional factors, such as the different levels of experience of amateur raters in crowdsourcing platforms or the role of social influence—which may differ between amateurs and professional raters.

B. The impact of global warming on subjective ratings

Our finding that higher temperatures in September are associated with lower ratings among both amateurs and professionals runs counter to previous findings for the northern hemisphere of a positive relationship between heat and wine quality (Jones and Davis, 2000; Ashenfelter, 2008b; Ashenfelter and Storchmann, 2010b). However, most of the previous analyses have focused on time frames spanning several decades. For example, Ashenfelter’s analyses of the Bordeaux equation cover the period between 1952 and 1980. Average temperatures have been steadily rising worldwide over the past decades, and Bordeaux is no exception. Figure 4 tracks the evolution of average temperatures in Bordeaux/Merignac over the past 70 years. The average temperature has

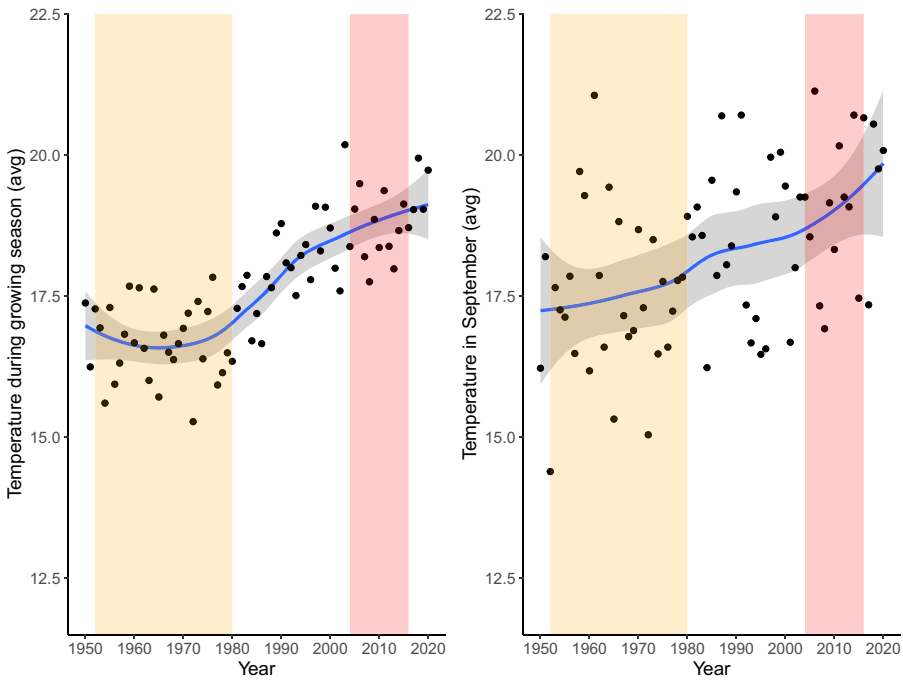


Figure 4. Evolution of average temperatures in Bordeaux.

Notes: Left panel: Average temperatures during the growing season (March–October). Right panel: Average temperatures in September. Source: <http://infoclimat.fr>, weather station: Merignac. Highlighted in lighter color are the vintages examined in Ashenfelter (2008b), from 1952 to 1980. Highlighted in darker color are the vintages considered in the current analysis, from 2004 to 2016. Curves and confidence bands show robust LOESS curves (locally estimated scatterplot smoothing using re-descending M-estimator with Tukey's biweight function) and their 95% confidence band.

increased by 1.6°C since Ashenfelter's seminal analysis: from 12.5°C (between 1952 and 1980) to 14.1°C (between 2004 and 2016).⁸

Excessively high temperatures can be detrimental to the quality of grapes as they have been found to inhibit certain biochemical pathways or physiological processes essential for the production of quality grapes (Deloire et al., 2004). In extreme cases, such high temperatures can cause premature veraison, high grape mortality through abscission, enzyme inactivation, and partial or total failure of flavor ripening (Mullins et al., 1992). From this perspective, the negative coefficients for temperature in September can be interpreted as early evidence for the effects of increased temperatures due to global warming on wine quality. Although our analysis is restricted by the lifespan of the Vivino platform (since 2010), the longer such crowdsourcing platforms are in operation, the more light can be shed on the relation between weather, climate change, and wine appreciation.

⁸With data from 1950 to 2020, a regression of average temperature during the growing season onto a time trend variable (and a constant term) yields a coefficient of 0.043 ($p < 0.001$). The same regression using average temperature in September yields a similar coefficient of 0.032 ($p < 0.001$).

V. Conclusion

Inferring the quality of a wine before tasting it has always challenged buyers. Freely available wine apps like Vivino democratize the production and consumption of information in the wine world, by making ratings from millions of wine-loving amateurs instantly accessible to consumers. But, how valid are these crowdsourced ratings and what can we learn from their insights?

We addressed these questions by creating and analyzing a rich new dataset of online Vivino reviews for a portfolio of red wines from Bordeaux. We assessed the validity of crowdsourced ratings based on two criteria: their consensus with the ratings of professional wine critics and their responsiveness to weather conditions known to affect wine quality. To this end, we matched our dataset with the reviews of professional critics and appended to it meteorological information collected from a nearby weather station.

We showed that Vivino ratings are overall consistent with those from professional critics. Not only is there broad agreement in vintage-level assessments but, crucially, the ratings also correlate substantially at the level of individual wines. Moreover, the amateurs' response to weather variability is in line with that of professional wine critics. Nevertheless, the average correlation between Vivino and critics' ratings is slightly lower than the correlations among critics themselves. Our exploratory analysis suggested that this discrepancy can be explained at least partly by differences in scope: While amateurs focus on immediate pleasure, professionals consider the wine's potential once it has matured. An implication of this finding for a prospective wine consumer confronted with contradictory reviews from critics and amateurs is as follows: If their intention is to find a good bargain they can invest in for the future, then the critics' rating might be a better guide than crowdsourced ratings. If, on the other hand, they are invited to a dinner party that night, then the average online rating of an app like Vivino might be more likely to help them make an impression.

Regressing averaged Vivino ratings onto yearly weather conditions provided additional evidence for the validity of these crowdsourced ratings. Averaged ratings from both sources were responsive to weather conditions known to affect wine production. Our analysis also suggests that global warming may already be having a discernible negative effect on wine quality, with ratings from professionals and amateurs alike being negatively associated with higher temperatures late in the growing season. While it has been predicted that the positive relationship between higher temperatures and wine quality in the northern hemisphere is eventually bound to "backfire" due to climate change, this is—to the best of our knowledge—the first empirical evidence for this turn of the tide. Should this trend persist, wine producers will need to adapt their practices and, for example, delay pruning dates or choose later-ripening varieties (Bordeaux Wine Council, 2019). Such adaptations to the wine-making bio-economy would help to avoid more radical changes in the geography of wine production. Based on our analysis, they are also necessary to maintain the quality of the wines produced, as reflected in consumer and expert ratings.

Overall, our analysis suggests that crowdsourced wine ratings are both valid and useful. If we are on the precipice of a paradigm shift whereby decentralized, crowdsourced reviews complement or even replace those of seasoned critics, then our analysis suggests that the future is in good hands.

Acknowledgements. Orestis Kopsacheilis acknowledges the American Association of Wine Economists for their scholarship as well as Dan Burdea for useful comments. We thank Susannah Goss for editing the manuscript.

Funding statement. Pantelis P. Analytis was supported by a Sapere Aude research leader grant by the Independent Research Fund Denmark. Ophelia Deroy was supported by the NOMIS foundation (Diversity in Social Environments project) and a Co-Sense grant from the Volkswagen Foundation. Bahador Bahrami and Karthykeya Kaushik were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (rid-o project, grant 819040).

References

- Analytis, P. P., Barkoczi, D., and Herzog, S. M. (2018). Social learning strategies for matters of taste. *Nature Human Behaviour*, 2(6), 415–424. doi:10.1038/s41562-018-0343-2.
- Ashenfelter, O. (2008b). Predicting the quality and prices of Bordeaux wine. *The Economic Journal*, 118(529), F174–F184. doi:10.1111/j.1468-0297.2008.02148.x.
- Ashenfelter, O., and Storchmann, K. (2010a). Measuring the economic effect of global warming on viticulture using auction, retail, and wholesale prices. *Review of Industrial Organization*, 37, 51–64. doi:10.1007/s11151-010-9256-6.
- Ashenfelter, O., and Storchmann, K. (2010b). Using hedonic models of solar radiation and weather to assess the economic effect of climate change: The case of Mosel Valley vineyards. *The Review of Economics and Statistics*, 92(2), 333–349. doi:10.1162/rest.2010.11377.
- Ashton, R. H. (2012). Reliability and consensus of experienced wine judges: Expertise within and between? *Journal of Wine Economics*, 7(1), 70–87. doi:10.1017/jwe.2012.6.
- 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. doi:10.1017/jwe.2013.18.
- Asimov, E. (2021). A wine worth waiting for. *The New York Times*.
- Bazen, S., Cardebat, J.-M., and Dubois, M. (2023). The role of customer and expert ratings in a hedonic analysis of French red wine prices: From gurus to geeks? *Applied Economics*, 56(1), 1–17.
- Bordeaux Wine Council (2019). Bordeaux in the face of climate change.
- Broomell, S. B., and Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, 74(3), 531–553. doi:10.1007/s11336-009-9118-z.
- Cardebat, J.-M., Figueat, J.-M., and Paroissien, E. (2014). Expert opinion and Bordeaux wine prices: An attempt to correct biases in subjective judgments. *Journal of Wine Economics*, 9(3), 282–303. doi:10.1017/jwe.2014.23.
- Chevalier, J. A., and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345–354. doi:10.1509/jmkr.43.3.345
- Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(1), 119–135. doi:10.1017/S0140525X00065675.
- Cicchetti, D. V. (2004). On designing experiments and analysing data to assess the reliability and accuracy of blind wine tastings. *Journal of Wine Research*, 15(3), 221–226. doi:10.1080/09571260500109368.
- Deloire, A., Carbonneau, A., Wang, Z., and Ojeda, H. (2004). Vine and water: A short review. *OENO One*, 38(1), 1–13. doi:10.20870/oeno-one.2004.38.1.932.
- Drichoutis, A. C., Klonaris, S., and Papoutsis, G. S. (2017). Do good things come in small packages? Bottle size effects on willingness to pay for pomegranate wine and grape wine. *Journal of Wine Economics*, 12(1), 104. doi:10.1017/jwe.2017.3.
- Galesic, M., Barkoczi, D., and Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, 5(1), 1–15. doi:10.1037/dec0000059.
- The Global Wine Score. (2017). Do the en primeur scores predict well the future quality of wines? *Medium*.
- Gokcekus, O., Hewstone, M., and Cakal, H. (2015). In vino veritas? Social influence on “private” wine evaluations at a wine social networking site. In Ashenfelter O, Gergaud O, Storchmann K, Ziemba W (Eds.),

- World Scientific Reference on Handbook of the Economics of Wine: Volume2: Reputation, Regulation, and Market Organization*, 423–437. World Scientific.
- Goldstein, R., Almenberg, J., Dreber, A., Emerson, J. W., Herschkwitsch, A., and Katz, J. (2008). Do more expensive wines taste better? Evidence from a large sample of blind tastings. *Journal of Wine Economics*, 3(1), 1–9. doi:10.1017/S1931436100000523.
- Goode, J. (2005). *The Science of Wine: From Vine to Glass*. University of California Press.
- Hodgson, R. T. (2008). An examination of judge reliability at a major US wine competition. *Journal of Wine Economics*, 3(2), 105–113. doi:10.1017/S1931436100001152.
- Holbrook, M. B. (1999). Popular appeal versus expert judgments of motion pictures. *Journal of Consumer Research*, 26(2), 144–155. doi:10.1086/209556.
- Jame, R., Johnston, R., Markov, S., and Wolfe, M. C. (2016). The value of crowdsourced earnings forecasts. *Journal of Accounting Research*, 54(4), 1077–1110. doi:10.1111/1475-679X.12121.
- Jones, G. V., and Davis, R. E. (2000). Climate influences on grapevine phenology, grape composition, and wine production and quality for Bordeaux, France. *American Journal of Enology and Viticulture*, 51(3), 249–261. doi:10.5344/ajev.2000.51.3.249.
- Jones, G. V., White, M. A., Cooper, O. R., and Storchmann, K. (2005). Climate change and global wine quality. *Climatic Change*, 73(3), 319–343. doi:10.1007/s10584-005-4704-2.
- Kurvers, R. H., Nuzzolese, A. G., Russo, A., Barabucci, G., Herzog, S. M., and Trianni, V. (2023). Automating hybrid collective intelligence in open-ended medical diagnostics. *Proceedings of the National Academy of Sciences*, 120(34), e2221473120.
- Le Mens, G., Kovács, B., Avrahami, J., and Kareev, Y. (2018). How endogenous crowd formation undermines the wisdom of the crowd in online ratings. *Psychological Science*, 29(9), 1475–1490. doi:10.1177/0956797618775080.
- Luca, M., and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12), 3412–3427. doi:10.1287/mnsc.2015.2304.
- Müller-Trede, J., Choshen-Hillel, S., Barneron, M., and Yaniv, I. (2018). The wisdom of crowds in matters of taste. *Management Science*, 64(4), 1779–1803. doi:10.1287/mnsc.2016.2660.
- Masset, P., Weisskopf, J.-P., and Cossutta, M. (2015). Wine tasters, ratings, and en primeur prices. *Journal of Wine Economics*, 10(1), 75–107. doi:10.1017/jwe.2015.1.
- Matthews, M., Anderson, M., and Schult, H. (1987). Phenologic and growth responses to early and late season. *Vitis*, 26, 147–160.
- McAuley, J. J., and Leskovec, J. (2013). From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In Schwabe D, Almeida V, Glaser H (eds), *Proceedings of the 22nd International Conference on World Wide Web*, 897–908. ACM.
- Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, 29(1), 1–16. doi:10.1016/j.jbusvent.2013.06.005.
- Muchnik, L., Aral, S., and Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science*, 341(6146), 647–651. doi:10.1126/science.1240466.
- Mullins, M. G., Bouquet, A., and Williams, L. E. (1992). *Biology of the Grapevine*. Cambridge University Press.
- Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, 78(2), 311–329. doi:10.1086/259630.
- Oczkowski, E., and Doucouliagos, H. (2015). Wine prices and quality ratings: A meta-regression analysis. *American Journal of Agricultural Economics*, 97(1), 103–121. doi:10.1093/ajae/aau057.
- Oczkowski, E., and Pawsey, N. (2019). Community and expert wine ratings and prices. *Economic Papers*, 38(1), 27–40. doi:10.1111/1759-3441.12240.
- Poni, S., Lakso, A., Turner, J., and Melious, R. (1993). The effects of pre- and post-veraison water stress on growth and physiology of potted Pinot Noir grapevines at varying crop levels. *Vitis*, 32(4), 207–214.
- Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532–535. doi:10.1038/nature21054.
- Storchmann, K. (2012). Wine economics. *Journal of Wine Economics*, 7(1), 1–33. doi:10.1017/jwe.2012.8.
- Stuenkel, 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. doi:10.1017/jwe.2015.3.
- Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Weiss, D. J., and Shanteau, J. (2004). The vice of consensus and the virtue of consistency. In *Psychological Investigations of Competent Decision Making* Smith, K., and Shanteau, J., Johnson, P. Cambridge: Cambridge University Press, 226–240.

A. Appendix

A.1 Overlap of reviews from Vivino and professional critics in matched dataset

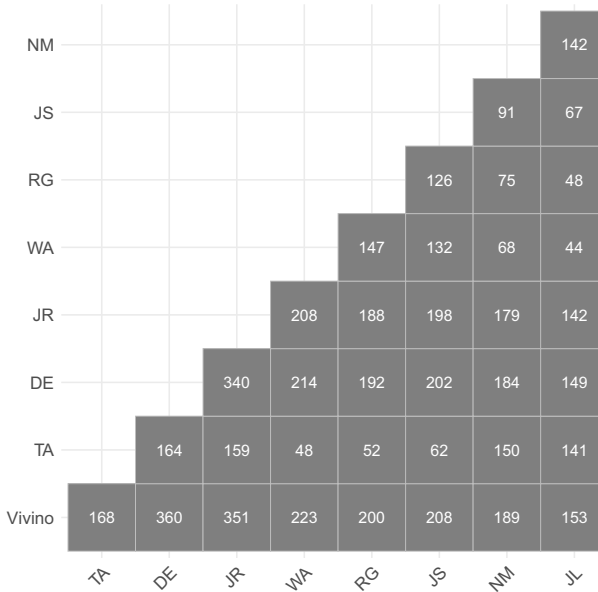


Figure A1. Number of matched wine reviews for Vivino and critics.

Notes: Each cell reports the number of wines from our initial portfolio of Bordeaux that were reviewed both in Vivino and by a professional critic. Vivino: averaged ratings from Vivino users. DE: Decanter; JS: James Suckling; JR: Jancis Robinson; JL: Jeff Leve; NM: Neal Martin; RG: Rene Gabriel; TA: Tim Atkin; WA: the Wine Advocate.

A.2 Summary statistics

Table A1. Summary statistics of variables used in ordinary least squares regression analysis

Statistic	Mean	St. Dev.	Min	Max
Avg. temperature in growing season	18.66	0.51	17.75	19.49
Rainfall in August	50.89	31.94	11.30	89.70
Rainfall in preseason	482.09	92.96	327.50	617.10
Avg. temperature in September	19.09	1.29	16.92	21.13
Vivino avg. ratings	3.96	0.24	3.00	4.64
Vivino avg. ratings (Z-scores)	0.00	1.00	-3.98	2.82
Prof. critics avg. ratings (Z-scores)	0.00	1.00	-2.30	3.17

Note: Average temperatures are measured in °C. Rainfall is measured in mm of water accumulated over the entire period. Vivino ratings are calculated by averaging at the level of the wine (i.e., the label of a chateau in a given vintage) and then transforming these averages into Z-scores by subtracting the mean and dividing by the standard deviation. We provide information about both levels here: before and after normal standardization. For the professional critics ratings, the process is the same but we add an additional step. Namely, we start by calculating Z-scores for each critic's ratings before averaging those Z-scores over the dataset. The reason for this additional step is that each critic uses their own rating system and it would be impossible to aggregate otherwise. Therefore, we only provide summary statistics for their Z-transformation here. The portfolio of wines for which these Z-scores are calculated is common for Vivino amateurs and professional critics and consists of 341 observations. The vintages we observe range from 2004 to 2016. The weather conditions we track follow the same 13-year period. The time trend we include in Table 1 is an index variable tracking these years.

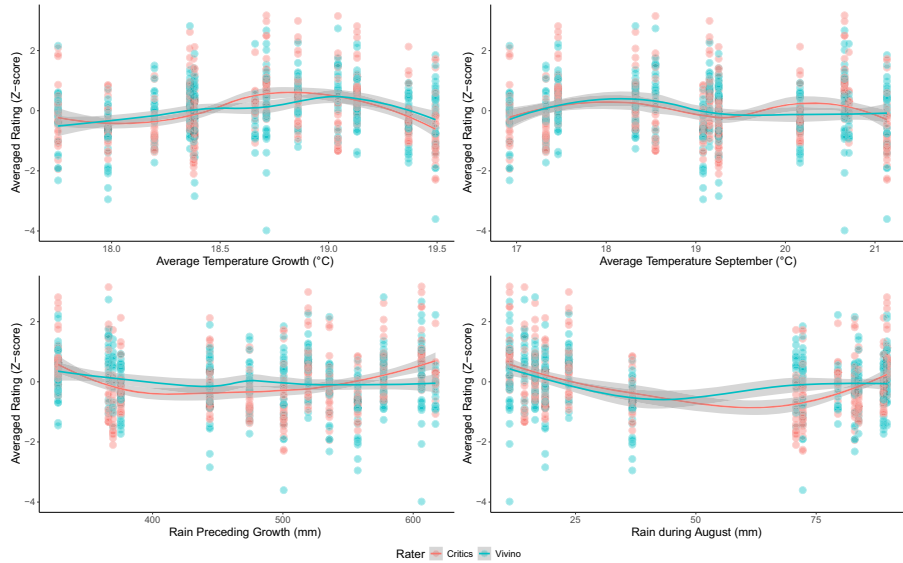


Figure A2. Averaged Vivino and professional critics' ratings (Z-scores) as a function of weather conditions. Note: Curves and confidence bands show robust LOESS curves (locally estimated scatterplot smoothing using re-descending M-estimator with Tukey's biweight function) and their 95% confidence band.

A.3 Alternative consideration of the “Bordeaux equation”

Table A2 reports on a regression analysis similar to that in Table 1, but specifically focusing on Vivino ratings only. Since matching Vivino ratings with those from professional critics is not necessary for this analysis, we can compare the coefficients of the unmatched dataset with those of the matched dataset. In the former dataset we observe all wine labels from our portfolio of red wines from Bordeaux across all years within our observation window (2004–2016). This results in a balanced panel data for model (1). In the latter dataset, there are gaps for some wine labels in certain years where we were unable to find at least three professional critics’ reviews. Reassuringly, the coefficients of both models are in high agreement, indicating that the matched dataset we used in our main analysis is unlikely to be significantly affected by selection effects.

Moreover, the regression model in Table A2 includes a variation of the set of independent variables compared to Table 1. The most notable difference is the addition of the variable “Average Age of wine” which is absent in Table 1 due to its inability to be calculated for professional critics’ ratings (critics give their ratings en primeur, when the wine has not yet been bottled and, therefore, has not aged). Age is computed as the difference between the review year and the wine’s vintage. It is then averaged across all reviews for a given wine (mean = 7.08, St.Dev = 2.83, min = 2.00, max = 13.94).

Additionally, unlike the model in Table 1, here we incorporate September in the calculation of the average temperature during the growing season and exclude the quadratic term for the average temperature in

Table A2. Determinants of wine ratings: Regressing Vivino ratings onto weather variables and average age

	(1) Vivino Unmatched	(2) Vivino Matched
Average temperature in growing season (April–September)	0.5487** (0.1598)	0.7061** (0.1567)
Average temperature in September	−0.1784*** (0.0368)	−0.2364*** (0.0398)
Rainfall in preseason (October–March)	0.0012 (0.0005)	0.0011 (0.0005)
Rainfall in August	−0.0043*** (0.0008)	−0.0067*** (0.0006)
Average age of wine	0.1276** (0.0353)	0.1833** (0.0657)
Time Trend	0.0685** (0.0245)	0.0988 (0.0452)
Observations	780	371
Number of groups	60	41
F-statistic	22.84	68.36
R ² for within model	0.3168	0.4089

Notes: The averaged, standard-normalized rating (Z-score) of a wine reviewed in Vivino is regressed onto climate variables (centered to their mean), the wine’s averaged age at the time of consumption and a time-trend. (1): Ratings from Vivino amateurs over entire portfolio of Bordeaux wines (balanced panel). (2): Ratings from Vivino amateurs over matched-with-experts portfolio of Bordeaux wines (unbalanced panel). The regression models include fixed effects at the level of the chateau and weights proportional to the number of ratings included in the calculation of each average. Weather variables have been centered to their mean. Each wine in our data set is uniquely identified by a chateau and a vintage. Average temperatures are measured in °C. Rainfall is measured in mm of water accumulated over the entire period.

Driscoll–Kraay (vintage- and chateau-clustered) standard errors are in parentheses.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

growing season. These adjustments are motivated by aligning more closely with the original formulation of the Bordeaux equation (Ashenfelter, 2008b). Despite these modifications, we observe strong agreement between the coefficients reported here and those in Table 1, indicating robustness in our conclusions regarding the factors influencing Vivino ratings across different variants of the Bordeaux equation.