Finding the Wise and the Wisdom in a Crowd: Estimating Underlying Qualities of Reviewers and Items

Nicolas Carayol and Matthew O. Jackson *

Draft: February 2024

Abstract

Consumers, businesses, and organizations rely on others' ratings of items when making choices. However, individual reviewers vary in their accuracy and some are biased – either systematically over- or under-rating items relative to others' tastes, or even deliberately distorting a rating. We describe how to process ratings by a group of reviewers over a set of items and evaluate the individual reviewers' accuracies and biases, in a way that yields unbiased and consistent estimates of the items' true qualities. We provide Monte Carlo simulations that showcase the added value of our technique even with small data sets, and we show that this improvement increases as the number of items increases. Revisiting the famous 1976 wine tasting that compared Californian and Bordeaux wines, accounting for the substantial variation in reviewers' biases and accuracies results in a ranking that differs from the original average rating. We also illustrate the power of this methodology with an application to more than forty-five thousand ratings of "en primeur" Bordeaux fine wines by expert critics. Those data show that our estimated wine qualities significantly predict prices when controlling for prominent experts' ratings and numerous fixed effects. We also find that the elasticity of a wine price in an expert's ratings increases with that expert's accuracy.

JEL Classification Codes: D80, C49, L66

Keywords: Ratings, Reviews, Qualities, Scoring, Aggregating Ratings, Wine Ratings

*Nicolas Carayol: Bordeaux School of Economics, UMR CNRS 6030, University of Bordeaux, Pessac 33608 FRANCE. Matthew O. Jackson: Department of Economics, Stanford University, Stanford, California 94305-6072 USA. Jackson is also an external faculty member at the Santa Fe Institute, and a fellow of CIFAR. Carayol owns a winery name Domaine de Cabrol (which is not in the Bordeaux region). Emails: nicolas.carayol@u-bordeaux.fr and jacksonm@stanford.edu. We thank Guillaume Forcade for giving us the expert wine ratings and price data, Orley Ashenfelter for a helpful discussion as well as conference and seminar participants in Paris, Melbourne, Bordeaux and the ASSA meetings. Nicolas Carayol also thanks the support of the USA-France Fulbright Commission and of the Bordeaux IdEx (Grant no ANR-10-IDEX-03-02).

1 Introduction

Most goods and services that humans consume are rated, including films, theater, art, books, wines, restaurants, stocks, scientific proposals, articles, and most consumer products. Platforms have led to enormous growth in the number of items that are evaluated and the number of people doing the rating. Some enterprises report consumers' ratings, while others collect evaluations from distributed sources and report them in one location. These ratings can come from experts (movie critiques' ratings) or from users (e.g. Yelp). Given that market efficiency improves when products' and services' qualities are better assessed (Akerlof, 1970), platforms, other market designers, and consumers can make use of ratings by exploiting the so-called "wisdom of the crowd," famously illustrated by Galton (1907).[1]

Using rating data efficiently is challenging, however, since many items have only a few ratings. For example, on *Amazon.com*—perhaps the largest and widest information aggregator on consumers goods in history with 182 million verified ratings of 12.1 millions rated products—80 percent of the products have fewer than 9 ratings, and 90 percent have fewer than 22 ratings.[2] Therefore, simply averaging ratings yields noisy and potentially biased estimates for the vast majority of items, given how few ratings they have.

We develop an estimation technique that gives substantial improvements over simple averages of ratings. We estimate people's biases and accuracies and then re-weight by accuracies after adjusting to compensate for the biases. This approach significantly improves item quality estimates because direct averaging under-weights the ratings of people who are discerning and over-weights others who are frivolous. An average is also susceptible to biases, as some people are consistently excessively negative (or positive) compared to the typical view of a product. In addition, some people only rate items with which they have extreme experiences, leading to a selection bias where they post excessively extreme ratings (e.g., see Nei (2017)).

The key idea is to use a reviewer's full set of ratings *across items* to evaluate their biases and accuracies. Even though there are few ratings per item, many reviewers typically rate multiple items. For instance, regarding the Amazon data referred to above, 36% percent of reviewers (28 millions distinct people) rate more than two items in the same product category, and those rate 4.57 items on average.

Estimating reviewers' biases and accuracies together with item qualities presents a chicken-and-egg problem: one needs some estimate of item qualities to estimate reviewers' accuracies and biases, and vice versa. We co-estimate these three things in a consistent manner. Our identification makes use of a reviewer's ratings on other items to discern the reviewer's bias and accuracy, which we then use to debias and appropriately weight their rating on any given item. Specifically, we show how this can be done via a slight variation of weighted

[1] For an example of some of the impact of reviews, see Chevalier and Mayzlin (2006); Tadelis (2016).

[2] In fact, out of the 21 product categories at *Amazon.com*, 19 are such that more than two-thirds of the items have no more than 10 ratings. For details on Amazon data, see Table A.1 and Figure A.1 in the Online Appendix A.

two-stage regression methods.

While we refer to 'true qualities' throughout, we emphasize that tastes are subjective. What we mean by 'true quality' is the anchor that would emerge if an infinite number of people rated the items on a common scale. When we refer to a reviewer's 'accuracy', we mean the extent to which their ratings match that average subjective value. Thus, having a high accuracy means that a given reviewer's ratings are good predictors of what many people's ratings would eventually converge to—after adjusting for each person's systematic bias. A reviewer with a low accuracy might still have "good taste" in some other sense, but is not as useful in predicting the consensus rating.

We show that our estimates of item quality, reviewers' biases and accuracies are unbiased and consistent. Using Monte Carlo simulations, we quantify the extent to which our method provides more accurate estimates of true underlying qualities than average ratings. The improvement is substantial (50% more accurate) even with a limited number of items being rated. Moreover, having more items enables us to more accurately estimate reviewers' biases and accuracies, which in turn improves our estimates of the qualities. We also show that the estimation gain is larger when reviewers are more biased, when they have more variance in their ratings, and have more heterogeneous accuracy.

We also illustrate our approach using 'expert' wine ratings. Fine wine ratings and markets constitute a relevant domain as i) there are large informational problems on wine/vintage quality (Ashenfelter, Ashmore and Lalonde, 1995; Ashenfelter, 2008), ii) though experts' opinions correlate, divergence among them is also frequent (Ashton, 2012; Hodgson and Cao, 2014), iii) ratings and prices are related (Dubois and Nauges, 2010; Friberg and Gronqvist, 2012; Hilger, Rafert and Villas-Boas, 2011) in ways that our approach sheds additional light upon, iv) there are large numbers of items that are reviewed by a relatively small number of prominent experts and it is interesting to identify their individual accuracies, and v) experts rate quasi-simultaneously and independently each wine vintage before bottling at the "en primeur" stage, thereby minimizing inter-expert influence.

This dataset includes a comprehensive set of wines, as well as the reviewers' identities and posted prices in retail outlets in three major markets (Paris, New York, and Hong Kong). Thus, we use our quality estimates to analyze the extent to which wine prices reflect underlying qualities and adjust to ratings. We find that our index is a significant predictor of wine prices, with a ten percent increase in our rating corresponding to a more than thirty percent increase in price, which is consistent with quality variations having large impacts on prices in this market. In addition, our index remains significant when accounting for prominent expert ratings (Parker and Robinson) and the highest rating (since many retail outlets selectively quote the highest ratings and/or most famous experts), as well as the variance in ratings (since consumers may be puzzled by inconsistent reviews). It also significantly predicts prices that emerge in the years that follow after harvesting, beyond what is learned from the average rating. Also telling is that we find that the elasticity of a wine price in an expert's ratings increases with our estimate of that expert's accuracy.

In other words, the reviews of wine experts whom we estimate to be more accurate, are better predictors of retail prices. Consistently estimating reviewers' accuracies is key to the performance of our method, and this provides supplementary external validity. In an Online Appendix, we also use re-rating data—experts often re-evaluate the exact same vintage of the same wines at later dates—to show that the adjustment in ratings is strongly predicted by our estimated quality, controlling for many other factors and fixed effects.

Our model also admits reviewers whose biases differ across different types of categories or types of items. For example, in our Bordeaux wine application, we examine the extent to which reviewers have different bias (tastes) and even different accuracies with respect to left- versus right-bank red wines (the two types of red wines that are produced in those sub-regions of the Bordeaux area, which use different grape combinations and have distinct features).

Although we apply the approach to wine reviews, the set of potential applications is enormous, as there are many other settings where people rate multiple items such as managers rating workers in firms, judges rating participants in athletic or artistic contests, referees reviewing articles or proposals for funding, critics reviewing new films or other media, customers offering feedback to companies from purchase experiences, students rating class instruction, and so forth.

Relation to the Literature

The idea of combining multiple opinions has been discussed since Condorcet (1785), and has been a topic of importance following Arrow (1951). This literature now spans from information processing from multiple sources (e.g., see Budescu, 2005) to social learning and herding (e.g., see Banerjee, 1992; Bikhchandani, Hirshleifer and Welch, 1992). Most recent papers focus on situations in which ratings arrive sequentially, and are interested in how previous reviews can influence, at least temporarily, subsequent reviews (e.g., see Muchnik, Aral and Taylor, 2013; Godes and Silva, 2012; Nagle and Riedl, 2014; Fradkin, Grewal and Holtz, 2018; Dai et al., 2018; Acemoglu et al., 2022). Such approaches are in order when dealing with rating data subject to social influence (e.g., in Dai et al., 2018). Our focus is instead on correcting for individual biases and inaccuracies and is well designed for the applications in which social influence is less of an issue, such as the *en primeur* wine reviews that we study, as well as the many others mentioned in the previous paragraph.

There have been previous studies, such as that by Budescu and Chen (2015), that show that forecasters' past records—e.g., the correctness of their past forecasts of market movements—can be used to identify better and worse forecasters, who can then be weighted to improve forecasts. Unlike forecasts which can be evaluated by examining the actual outcome, we never see the underlying true qualities, and so those have to be inferred. Our innovation in the formulation of the problem and the identification that allows us to estimate the qualities, biases, and accuracies in a consistent, unbiased manner.

Our method can provide immunity to some types of manipulation of ratings, as well

as selection biases in ratings.[3] If a reviewer only gives high ratings, then that bias can be identified. If a reviewer only gives ratings when they have extreme experiences, then they are more likely to estimate when they are making large errors and their accuracy will suffer. Both of these sorts of systematic deviations are identified by our technique.[4]

Our analysis may also provide a rationale for collecting datasets of ratings. Firstly, as our approach is shown to perform well when there are relatively few ratings per item, it suggests collecting ratings over different items from the same reviewers, who can then be weighted appropriately. Secondly, acknowledging that once social influence is in the data, undoing it may be challenging, our approach suggests soliciting independent evaluations. This may be at the expense of reducing sample size, but then, thanks to our method, one can still obtain reliable item evaluations with significantly less, but well chosen, data points.[5]

Our work could also serve as a foundation for an analysis that deals with social interference.[6] The closest paper in that regard is Dai et al. (2018), but the additional structure in their analysis (having classes of reviewers, biases and accuracies that are class dependent) avoids the identification problems we face. One could extend/merge our and their analyses, to address setting where ratings are subject to nontrivial peer influence while allowing for individual biases and accuracies.

Our method is easy to apply. It can be viewed as a variation of two-stage weighted least squares regressions that allows for groupwise heteroscedasticity (e.g., Greene, 2010) as well as two-way (item and reviewer) fixed effects.[7] A "group" in our context is the set of ratings by a particular reviewer and different reviewers can have different variances in their error terms. However, our approach differs from the standard groupwise heteroscedasticity models in three ways. The first is that the items in our model are the objects of interest. Usually heteroscedasticity is an issue that needs to be addressed for efficiency. Here, variances of errors are the accuracies of the reviewers and are of specific interest, as our method can thus be used to evaluate reviewers and not just items. Second, the addition of item fixed

[3]For some of the literature on incentives in rating and recommender systems see Resnick and Zeckhauser (2002); Dellarocas (2003); Resnick and Sami (2007); Ekstrand, Riedl and Konstan (2011); Ricci et al. (2011).

[4]It is impossible to completely eliminate some forms of manipulation. For example, if a reviewer who has a history of accurate ratings can manipulate a single item's rating in a case when there are few other ratings. However, if someone gives high ratings to some subset of items, then that is more easily identified. Also, our system takes into account the history of a reviewer, and so the ratings of shills are largely ignored.

[5]For instance, a platform selling goods may experimentally generate a limited number of rating data points free of social influence (previous ratings are not observable). The platform may then rely on our method, applying it to the experimental data, as we show it performs well even with relatively few ratings per item.

[6]There can be a variety of different forces and incentives at work (conformity, enmity, counteracting trends, selection, herding) which makes it a context-specific problem, which may also explain why the dynamic interactive effects may be small in some cases (e.g., see Figure 1 in Askalidis and Malthouse, 2016). Thus, we leave it for future research.

[7]One has to be careful to use the appropriate normalization, as discussed below, so that reviewer biases sum to some known constant (usually zero), but allowing the item fixed effects which are the item true qualities to be correctly estimated. Without this normalization, one would obtain a translation of our estimates.

effects (the true qualities) and reviewer fixed effects (reviewer biases) to such an analysis, together with the fact that each review involves one item and one reviewer, leads to an identification problem. By adding a constant to all reviewer fixed effects and subtracting it from all item fixed effects one ends up with the same equations. Solving this requires normalizing the reviewer fixed effects to sum to some constant (usually, but not always, zero) and then introduces a constraint. Many two-way fixed effects regressions run into the same issue and so some normalization is built into most software. Here it is essential in terms of interpretation that the normalization be in terms of the biases so that one can then estimate and correctly interpret the item fixed effects, which are the estimated true item qualities. With other normalizations, one would obtain translations of our estimates. Third, the structure of this problem is one where one cannot iterate on the number of stages of estimation beyond the second stage as in some other heteroscedasticity estimations (e.g., Carroll and Ruppert, 1982). Again, this is an identification issue that arises from the structure of our setting, this time from the fact that each reviewer only provides one rating of each item, as we discuss more fully below.

Our paper also contributes to the literature on expert ratings in the wine market. Ashenfelter, Ashmore and Lalonde (1995) initially expressed doubts about the value of information contained in those ratings, and experts' opinions have been shown to diverge even within relatively homogeneous sub-segments of the market (Ashton, 2012; Hodgson and Cao, 2014). Cao and Stokes (2010) analyze ways in which experts' ratings may be inaccurate.[8] The fact that reviewers' idiosyncratic scales and biases can impact some aggregate score was first pointed out in the context of wine reviewers by Ashenfelter and Quandt (1999) (see also Lindley, 2006), when analyzing the results of a famous tasting in 1976 that put California wines on the map when 'winning' a tasting against a selection of top French wines. Ashenfelter and Quandt (1999) convert reviewers' scores into rankings that are then aggregated, which can work well when all reviewers rate the same set of objects, but does not work more generally.[9] The renormalization that we use—though different in spirit—generalizes such a ranking method since it accounts for biases and differences in accuracies, and also allows reviewers to be rating different sets of items.

The literature has also examined the relation between reviewers' ratings and prices. Ali, Lecocq and Visser (2008), Dubois and Nauges (2010), Friberg and Gronqvist (2012), and Hilger, Rafert and Villas-Boas (2011) find that well-known experts' ratings predict prices.[10] Rather than analyzing the impact of experts' judgements on the wine market, or their disagreements, we instead use those opinions to estimate latent wine quality. We also estimate

[8]Two of their terms 'bias' and 'variation' are superficially similar to what we term bias and accuracy, but they are distinct in the way that they are estimated as well as what they actually mean.

[9]Note that using rankings goes beyond normalizing scales so that they are comparable, and then averaging. See also the discussion in Ekstrand, Riedl and Konstan (2011), as well as the method of Gergaud, Ginsburgh and Moreno-Ternero (2021) for going beyond averaging.

[10]See also Cardebat, Figuet and Paroissien (2014), who conclude that fundamentals prevail in explaining wine prices when accounting for experts' residual subjectivity.

the correlation between estimated quality and prices conditional on lead experts' ratings, but the purpose of this is different. We are checking that our estimated wine qualities are indeed good predictors of prices, consistent with the idea that markets learn about quality.

We should also mention a growing literature in computer science about recommender systems (Ekstrand, Riedl and Konstan, 2011; Ricci et al., 2011). A goal of such techniques is to provide suggestions to consumers on the basis of their previous evaluations or choices, others' ratings, and products' characteristics. Our approach is different as we estimate underlying item quality and simultaneously uncover how accurate reviewers are, which complements the previous literature on recommender systems.

The paper is organized as follows. In Section 2, we introduce the setting, in Section 3, we present our approach, and we illustrate the ideas behind our approach on a famous case study: the Paris 1976 wine contest. In Section 4, we demonstrate the properties and added value of our estimation analytically and via Monte Carlo simulations. In Section 5, we apply the approach to a data set of wine experts' ratings of Bordeaux wines. In Section 6, we document the relation between our estimated qualities of Bordeaux wines and retail prices. In an appendix, we study experts' adjustments of ratings of wines they earlier rated as "en primeur".

2 Reviewers, Items and Ratings

2.1 Notation

Consider a set N of items $i = 1, ..., n$, that may be rated, and a set M of reviewers $j = 1, ..., m$, each rating a specific subset of the items $N_j \subset N$. We use the term reviewers as a generic term encompassing users, reviewers, critics, and experts. Reviewers are just a collection of people who rate items—some of whom may do it for a living while others only rate the occasional items that they consume.

Each reviewer rates an item at most once (the extension to more ratings is straightforward). Let 1_{ij} be the indicator variable that is 1 if reviewer j rated item i, and 0 otherwise. Let $m_i = \sum_j 1_{ij}$ be the number of the number of ratings of item i and $n_j = \sum_i 1_{ij}$ the number of ratings by reviewer j. The total number of ratings is given by $R = \sum_{ij} 1_{ij}$.

The ratings are listed in an $n \times m$ matrix g with the $g_{ij} \in \mathbb{R}$ being j's rating of item i, and with g_{ij} missing when j did not rate item i.

2.2 Reviewers' Ratings of Items

When rating an item i, a reviewer j (independently) estimates the unobserved true quality of that item q_i. Reviewer j may have a systematic *bias* b_j (for instance always over- or always under-rating items). Moreover, reviewers are not perfect and so each rating is likely

to include an error ε_{ij} on the top of the systematic reviewer-specific bias. A simple and natural way to take those three dimensions (true quality, bias and error) into account is to let reviewer i's observed rating be defined by

$$g_{ij} = q_i + b_j + \varepsilon_{ij}, \tag{1}$$

where the errors ε_{ij} are centered at 0, independent across js and i.i.d. for the same j, with standard deviation σ_j, since the bias term picks up systematic over or under-rating.

A measure of the "*accuracy*" of reviewer j is $a_j \equiv \frac{1}{\sigma_j^2}$: the inverse of her squared error. This corresponds to the standard definition of statistical precision.

It does not matter to our analysis how people choose which items they rate as long as their ratings satisfy equation (1) for the items that they rate. For instance, it is ok if a wine reviewer only chooses to rate wines that sell large quantities, or small quantities, or have received high past ratings, or for which the prior expectation of q_i is high, as long as the current rating still has an independent error term associated with it as in (1).

2.3 "Quality," Subjective Tastes, and Categories

An assumption is that there is some 'true' underlying quality q_i. In cases in which people are really assessing some objective quantity, like the weight of an ox in Galton (1907), there is a an objective reality. Instead, in most applications that we have in mind people are assessing items that are multidimensional and *subjective*, like a wine, movie, restaurant, art, or other good or service. It might not be that all people would ever agree on that quality, even with enormous experience.

What we mean by "true quality" is the average rating if an infinite number of unbiased people all rated this item. This is still a useful exercise since people have correlated tastes and knowing this answer can help people predict how much they would personally enjoy the item. Thus, even though we use the term "true quality," this should be interpreted as an "anchor," around which people may disagree, but when they disagree it comes from two sources: a systematic individual bias in taste and a term which is random from our perspective. For instance, if we examine the ratings of a set of romantic comedy movies, it might be that some particular reviewer tends to like this genre more than other reviewers, providing relatively inflated ratings, and thus has a positive bias. Even reviewers for whom we have only a few ratings still provide useful information.

Although we use the term "bias"—matching statistical terminology—a biased reviewer's ratings are still useful since once we adjust for that bias, their *relative ratings* provide valuable information. An accurate reviewer with a large but well-estimated bias is more informative than an unbiased reviewer with a lower accuracy.

If there are multiple categories of items, then reviewers' biases may differ from one category to the other. We could add item characteristics and reviewer specific coefficients on the

right side of Equation 1, substituting for b_j with $b_{ij} = b_j + \beta_j X_i$ for instance. That would require that reviewers' accuracies not vary with item types, whereas it is likely that if reviewers are more or less appreciative of items depending on their categories, then they may also be more experienced and knowledgeable on some category of items than others. Therefore, we propose a generalization of our model in Section 5.5 that fully accounts for both different biases and accuracies across item types, and propose associated statistical tests. For the purpose of exposition, we relegate the discussion of dealing with different categories of items to Section 5.5.

3 Estimating Items' Qualities and Reviewers' Accuracies

3.1 "True Qualities" of the Items

Note that one can also write (1) as a linear set of equations of R observations with generic observation r described by

$$g_r = \sum_i q_i I_{ir} + \sum_j b_j I_{jr} + \varepsilon_r, \tag{2}$$

where I_{ir} is an indicator variable taking value 1 if and only if observation r is a rating of item i, I_{jr} is an indicator variable taking value 1 if and only if observation r is a rating by reviewer j, and ε_r has variance σ_j^2 where j is the reviewer associated with observation r.

Thus, if we knew the σ_j^2s then we could estimate the true qualities, q_i's, and biases, b_js, by weighted least squares. This would mean solving:

$$\min_{(q_i, b_j)_{ij}} \sum_r \left(\frac{g_r - \sum_i q_i I_{ir} + \sum_j b_j I_{jr}}{\sum_j I_{jr} \sigma_j} \right)^2. \tag{3}$$

The weighting is by the precision or accuracy of each reviewer, which also aligns with a Bayesian weighting under the model above.

There are two reasons that this cannot be done directly. One is that we do not know the σ_j^2s, and the other is that many data sets are such that a reviewer only provides at most one rating per item. This presents an identification problem that leads to collinearity and a rank deficiency in the independent variable matrix.

3.2 Identification and Other Estimation Challenges

To understand why the system is not fully identified, note that the first-order conditions that weighted least squares solutions must solve are

$$\widehat{q}_i = \frac{\sum_j \frac{1_{ij}\left(g_{ij} - \widehat{b}_j\right)}{\sigma_j^2}}{\sum_j \frac{1_{ij}}{\sigma_j^2}}, \tag{4}$$

$$\widehat{b}_j = \frac{\sum_i 1_{ij}\left(g_{ij} - \widehat{q}_i\right)}{n_j}, \tag{5}$$

for each i, j, using notation of g_{ij} instead of g_r since reviewers provide at most one rating per item. (4)–(5) form a system of $n+m$ equations in the same number of unknowns. Nonetheless, it is not well-identified: if we decrease all qualities by some constant c and increase all reviewers' biases by the same amount, then we still have a solution to these equations.[11] Another way to see this is to note that the matrix of independent variables has a deficient rank. Just as an example, for each pair of reviewers that estimate the same pair of items, the sum of the rows corresponding to reviewer 1 on item 1 and reviewer 2 on item 2 has a 1 in each of those two items and those two biases, and the same is true of the sum of the rows that assign reviewer 1 on item 2 and reviewer 2 on item 1. The rank of the matrix is generally less than $n+m$, and insufficient for the regression.

We need at least one additional equation to normalize the values of the biases and qualities and tie things down. In particular, a natural way to tie things down, is to require that biases are *on average* 0.[12] This then delivers an unbiased estimate of the true quality relative to the population of observed reviewers. Thus, we normalize the biases to have mean 0, via the constraint that

$$\sum_j \widehat{b}_j = 0. \tag{6}$$

Once this constraint is added, we can re-express $b_m = -\sum_{j<m} b_j$, which increases the relative rank of the matrix.

Note, however, this is not always sufficient for identification. In particular, suppose that we can partition the items into two disjoint subsets $N = N_1 \cup N_2$, and similarly for reviewers $M = M_1 \cup M_2$, such that reviewers in M_1 only rate items in N_1 and reviewers in M_2 only rate items in N_2. Then for any set of qualities and biases that are optimizers of the (weighted) least squares problem, one could also increase the biases of reviewers in M_1 and correspondingly decrease the value of the items in N_1; and then do the reverse for

[11] This is not just an issue from the way in which indicator variables have been specified. The same issue is true of equation of true underlying values (1), where one could offset biases and ratings.

[12] In fact any real could be used instead of 0. For instance if one has good reasons to believe that the set of reviewers is not representative and has further data on how the average bias in the group relates to the overall average bias, one may use this correction.

the reviewers in M_2 and items in N_2, in a way that they offset each other and still respect the overall constraint $b_m = -\sum_{j<m} b_j$ and all of the equations for the regression. Thus, we need to rule out such subsets. This then ensures that one can get a full cycle of items and reviewers - every item is rated by at least one reviewer, and we can find a cycle defined by overlaps that includes all reviewers such that reviewer 1 overlaps in the set of items with reviewer 2 who overlaps with reviewer 3, and so forth.[13]

If the matrix of independent variables is normalized to have biases to sum to zero and has sufficient rank (necessarily ruling out the partitions described above), then one can run the corresponding (restricted) variance-weighted least squares estimation problem, where the weights are unknown and assumed to be constant across each individual reviewer.

3.3 A Two-Stage Estimation

Once we have the appropriate normalization, and sufficient rank of the item-reviewers matrix we then run a heteroscedastic-consistent two-stage weighted least squares analysis.[14] In particular, in the first stage we run a standard regression with the normalized matrix, and from that estimation we obtain a set of residuals e_{ij}s. From these we estimate

$$(\sigma_j^{one})^2 = \sum\nolimits_i \frac{1_{ij} e_{ij}^2}{n_j}. \tag{7}$$

This imposes more structure than the usual heteroscedastic-consistent analysis, as we have the same variance for each rating by any given reviewer.

In the second stage we run a weighted regression with the normalized matrix and weights of

$$w_{ij} = w_j = \frac{\frac{1}{\left(\sigma_j^{one}\right)^2}}{\sum_{j'} \frac{1}{\left(\sigma_{j'}^{one}\right)^2}}, \tag{8}$$

where j is the reviewer on observation ij (or more generally, for the observation r if there are multiple ratings of the same item per reviewer). This leads to second stage estimates, q_i^{two} and b_j^{two}, that can be used to re-infer the variance in reviewers' errors:

$$\left(\sigma_j^{two}\right)^2 = \frac{\sum_i 1_{ij} \left(g_{ij} - b_j^{two} - q_i^{two}\right)^2}{n_j}. \tag{9}$$

.

[13] This plays more of a role in identification with smaller numbers of reviewers. With large numbers of reviewers, biases will naturally tend to average to be close to 0 across different subsets of reviewers.

[14] As an alternative to normalizing the biases to sum to 0 within the matrix, one can alternatively add it as a restriction on the regression and run a restricted (weighted) regression. As many software packages will not accompany both the restrictions and the two stage weighted process with the constraints on the error terms, it is often easier to do it the way we have.

We remark that if every reviewer rates every item, then the first stage has a simple solution of:

$$q_i^{one} = \sum_j \frac{1_{ij} g_{ij}}{m_i}, \tag{10}$$

and we can estimate the b_j's via

$$b_j^{one} = \sum_i \frac{1_{ij} \left(g_{ij} - q_i^{one}\right)}{n_j}. \tag{11}$$

Given these estimates, we then get an estimate for the error variances:

$$(\sigma_j^{one})^2 = \sum_i \frac{1_{ij} \left(g_{ij} - b_j^{one} - q_i^{one}\right)^2}{n_j}. \tag{12}$$

However, more generally, if some reviewers do not evaluate all items, then the first order conditions that characterize the first stage of the regression are more complicated, but correspond to the standard ones (given the normalized matrix of independent variables). What happens above is that each q_i is averaged across all reviewers, and so their biases all cancel out, greatly simplifying the estimation. When not all reviewers evaluate all items, then these no longer cancel and one has to solve the standard regression equations for the normalized matrix.

We point out that one cannot iterate any further on this procedure. In some settings with unknown heteroscedastic error variances, it is possible to continue to iterate, using estimated weights to re-estimate errors and then to iterate to convergence, as for instance described by Carroll and Ruppert (1982). Here, if we iteratively re-estimate the q_is with the latest iteration's weights based on the latest iteration's estimates of the errors σ_js, then the process can converge to setting some $\sigma_j = 0$, which becomes self-fulfilling. For instance, consider a setting in which some reviewer j^* rates each item. If we set that $\sigma_{j^*}^2$ to be 0, then that results in each quality $\widehat{q}_i$ estimate to be equal to g_{ij^*}, and the overall errors to be 0, justifying the estimated variance. By only using two stages, the second stage only uses the first stage and the weights have no chance to influence their own estimates. Once a third stage, or beyond is used, weights influence their own estimates, which can then bias those estimates. For instance, with more weight on the most accurate reviewers, one gets better fits by further increasing their relative accuracy.

3.4 An Example Using a Well-Known Application: The Paris 1976 Wine Contest

Before exploring the properties of our approach and its added value with respect to other methods, we illustrate how and why our methodology may lead to different estimates in the context of a well-known case study.

We re-examine the results of the famous and consequential 1976 wine tasting that in-

cluded red wines from both California and Bordeaux,[15] since others, including Lindley (2006) and Ashenfelter and Quandt (1999), have used it to discuss methods of recombining experts' ratings. In particular, the variance in the experts' ratings, as noted by Lindley (2006) for instance, has generated attention. In an important paper, Ashenfelter and Quandt (1999) (see also Hulkower, 2009) suggest that a way to combat the variance is to convert the scores into rankings by each expert, and then average the rankings rather than the raw scores. They point out that raw averages lead to undue influence of noisy reviewers, since an inflated score can distort an average score. Their solution of using rankings instead of actual ratings gives each expert an equal influence on the outcome. However, giving each expert an equal influence allows experts who are biased and inaccurate to have the same influence as those who are unbiased and significantly more accurate. This is where our method can improve.

The rankings obtained via the three approaches on the 1976 tasting are summarized in Table 1. The table lists our estimated qualities (q_i^{two}), the average ratings ($q_i^{avg} = \sum_j \frac{1_{ij} g_{ij}}{m_i}$, as used in the original contest), and the Borda scores (q_i^{Borda}, as in Ashenfelter and Quandt, 1999, and Hulkower, 2009). Wines are ordered according to our estimated quality. The Borda method still finds Stag's Leap as the winner, but results in some other shifting of the rankings as compared to the average rating. Our method leads to a ranking which differs from the other two, in particular in that Château Montrose ends up with the highest score, slightly edging out Stag's Leap.[16] Additionally, our estimates of experts' biases and accuracies reported in the bottom part of Table 1 show that although biases are relatively small, there is a large variation in experts' accuracies, which suggests that equal weighting as in averaging or Borda scoring, is inappropriate.

It is informative to discuss why Château Montrose beats Stag's Leap according to our technique's ranking whereas the two other methods lead to the reverse conclusion. The raw data show that two experts give their top ratings to the two wines: Raymond Oliver, who's the most accurate expert according to our estimates and Steven Spurrier. It is also a tie for Patricia Gallagher. Leaving aside those three experts, we are left with two groups of experts who have opposing views on those two wines. In the first group, Aubert de Villaine, Christian Millau, Jean-Claude Vrinat, Michel Dovaz and Pierre Tari, all rate Château Montrose above Stag's Leap, but by a small margin. In the second group, Odette Kahn, Christian Vanneque and Pierre Brejoux, all prefer Stag's Leap by a much larger margin (from 2 to 5.5 points). The Borda approach mitigates the influence of those three reviewers who provide extreme ratings by essentially giving each reviewer the same influence. Our approach goes further

[15] The 1976 tasting was famous because the highest average rating was given to Stag's Leap of Napa Valley, above some of the finest French wines, which resulted in widespread press coverage and helped establish the reputation of California as a producer of high-quality wines, not just bulk wines. We include only the red wines, as the data on the white wines have some issues, as discussed in Hulkower (2009).

[16] Gergaud, Ginsburgh and Moreno-Ternero (2022) trace ratings over many years for these wineries and by more reviewers, and find that the ratings of the wineries differ when other years and ratings are taken into account.

Table 1: Quality rankings of Cabernet red wines in the Paris 1976 contest according to our method, the arithmetic average and the Borda score, as well as our estimates of experts' biases and accuracies.

Wines and Vintages	q_i^{two}	q_i^{avg}	q_i^{Borda}
Château Montrose 1970 (F)	13.93	13.64	68.5
Stag's Leap Wine Cellar 1973 (CA)	13.89	14.14	69
Château Mouton Rothschild 1970 (F)	13.87	14.09	67
Château Haut Brion 1970 (F)	12.76	13.23	61
Ridge Monte Bello 1971 (CA)	11.98	12.14	55
Château Léoville-Las-Cases 1971 (F)	11.33	11.18	37.5
Heitz Martha's Vineyard 1970 (CA)	10.61	10.41	40
Mayacamas 1971 (CA)	10.36	9.77	32.5
Clos du Val 1972 (CA)	9.87	10.14	30.5
Freemark Abbey 1969 (CA)	9.77	9.64	34
Experts	$\left(\sigma_j^{two}\right)^2$	A_j^{two}	b_j^{two}
Aubert de Villaine (owner/manager, Domaine de la Romanée-Conti)	6,36	0,92	-0,84
Christian Vanneque (sommelier, restaurant La Tour D'Argent)	15,15	0,39	0,11
Claude Dubois-Millot (sales director, guide Gault et Millaud)	5,19	1,13	-0,24
Jean-Claude Vrinat (owner, restaurant Taillevent)	3,60	1,63	-0,14
Michel Dovaz (Institut du Vin)	6,08	0,96	-0,29
Odette Kahn (director, La Revue du Vin de France)	13,57	0,43	-2,64
Patricia Gallagher (l'Académie du Vin)	6,38	0,92	2,06
Pierre Brejoux (inspector general, Inst. Nat. des Appellations d'Origine)	6,96	0,84	0,16
Pierre Tari (owner, Chateau Giscours)	6,15	0,95	1,66
Raymond Oliver (owner and Chef, Restaurant Le Grand Vefour)	3,58	1,64	-0,24
Steven Spurrier (l'Académie du Vin)	4,97	1,18	0,36

Notes: the normalized accuracy is accuracy divided by average accuracy among experts: $A_j^{two} = \frac{\sum_{j'}\left(\sigma_{j'}^{two}\right)^2}{m\left(\sigma_j^{two}\right)^2}$.

because we identify the accuracy of each expert and weight them accordingly rather than equally. It turns out that the three experts who have strict preferences for Stag's Leap over Château Montrose are estimated to be the least reliable experts overall by estimating the variances in their ratings, so that their strong views in favor of Stag's Leap are beaten by the lower variance judgments of the experts from the other group who prefer Château Montrose.

Of course, the statements from Section 2.3 apply here about interpreting our "quality" estimate as an anchor around which a large population of people's *subjective* tastes will be distributed. So, having a higher quality simply means that, on average, people would rate this higher; but does not mean that one wine is "better" than another in some objective sense—just that this is the right mean in terms of predicting the overall population's ratings. Additionally, the quality estimates of the top three are nearly identical and statistically indistinguishable.

This example shows why our method can provide estimated qualities that differ from both the average rating and that derived from Borda. There are other studies that suggest other voting methods—in which Montrose or other French wines end up on top—such as: majority voting ranking (Balinski and Laraki, 2013), or finding a Condorcet winner or via various Shapley Value calculations (Ginsburgh and Zang, 2012). The fundamental difference between our approach and others, is that others consider different ways of combining scores but still treat all reviewers as equals. Instead, we are discriminating between reviewers based on estimated accuracies via a fixed point in which we estimate true underlying qualities. Since we have no known "truth" in this application with which to compare methods, it just provides some insight into the types of adjustments that our approach makes.

It is important to note that our approach is aimed at providing more accurate estimates of "quality" and that in this example the ranking is a by-product. With any finite sample, the rankings derived from point estimates can be incorrect, especially when there are nearly-tied scores, and so we do not interpret our ranking as being correct. If one wants to understand the potential errors in rankings, one can use techniques such those developed in Mogstad et al. (2020) to account for the finite sample noise.

4 Properties and Gains of the Two Stage Estimation

To provide more insight about how our quality estimates compare to the truth, we next explore the properties of our q_i^{two}, b_j^{two} and σ_j^{two} via analytic results and Monte Carlo simulations.

4.1 Consistency, Unbiasedness and Gain

Consider a sequence of ratings indexed by R from the data generating process described in Section 2. Let reviewers' biases be i.i.d. distributed with mean 0 and variance σ_b^2. Let the overall distribution of errors have variance σ_ε^2, and suppose that the distribution of variances

of reviewers has support $\left[\underline{\sigma}_{\varepsilon}^2, \overline{\sigma}_{\varepsilon}^2\right]$.[17] Let each reviewer review at least $n(R)$ items and each item be reviewed by at least $m(R)$ reviewers, where $n(R)$ and $m(R)$ both go to infinity as R grows, and suppose that the biases are normalized to sum to 0 and the rank condition is satisfied.

It then follows by standard arguments that the estimators q_i^{two} and b_j^{two}, are consistent (e.g., see White, 1980), and then given the growing $n(R)$, so are the estimated variance terms $\left(\sigma_j^{two}\right)^2$. We also provide conditions under which the estimates are unbiased.

LEMMA **1** *In addition, if the reviewers' biases and errors each have symmetric distributions around 0, then q_i^{two} and b_j^{two} are unbiased and distributed symmetric about their means.*

Consistency and unbiased results are reassuring, but we would also like to compare our estimates of item qualities q_i^{two} from the two stage procedure with the straight averages (q_i^{avg}), when $n(R)$ and/or $m(R)$ remain relatively small.

Note first that squared error of the simple average estimator is simply

$$E\left[(q_i^{avg} - q_i)^2\right] = \frac{\sigma_b^2}{m(R)} + \frac{\sigma_{\varepsilon}^2}{m(R)}. \tag{13}$$

Estimation errors come from two sources, reviewers' biases and their errors, which are both moderated by the number of evaluations per item.

When we instead use our approach, then the biases are at least partly eliminated which reduces the first term (to zero as $n(R)$ grows, even with a small $m(R)$). The second term is also reduced, since the largest variances are reduced with lower weightings - so instead of a straight average of errors, larger variance errors receive lower weights.

4.2 Monte Carlo Simulations

To see how efficiently our method estimates true qualities as a function of the sample size and how it compares to other methods, we perform Monte Carlo simulations in which we know the true qualities and can then directly compare different methods.

Item qualities are randomly drawn from a uniform distribution $q_i \sim \mathrm{U}\left(\underline{q}, \overline{q}\right)$. Reviewers' biases are randomly generated from a centered normal distribution $b_j \sim \Phi\left(0, \sigma_b^2\right)$, and their mean errors are drawn according to uniform distribution $\sigma_j \sim \mathrm{U}\left(\underline{\sigma}, \overline{\sigma}\right)$. Given items' qualities and reviewers' biases and accuracies, ratings are generated according to equation (1). Since in many applications not all reviewers rate all items, only a random proportion f of the cells of the $n \times m$ matrix g are filled, and the remaining cells are left empty.

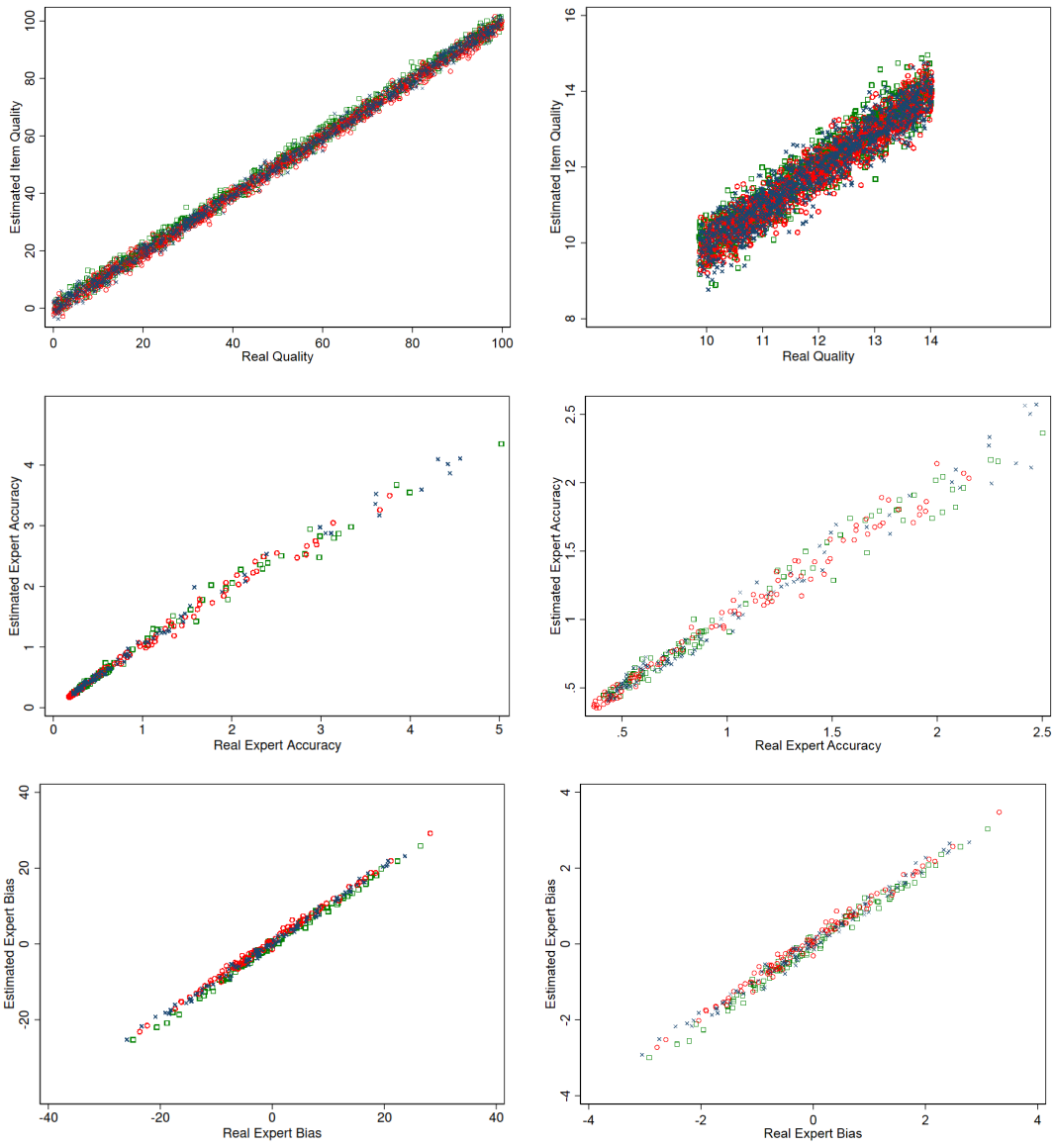
We first compare our estimates of qualities to the true values with two different sets of simulations. In one, we use 100 reviewers and 1,000 items and $f = .5$. True qualities are drawn uniform on $[0, 100]$, with a standard deviation on biases of $\sigma_b = 10$ and reviewers'

[17] If a variance estimate $(\sigma_j^{one})^2$ and $(\sigma_j^{two})^2$ lands outside of these bounds, reset it to the closest endpoint.

standard deviations drawn between $\underline{\sigma} = 5$, and $\overline{\sigma} = 25$. In the other set we set things based on the 1976 red wine tasting contest (see Section 3.4).[18]

Figure 1 plots our estimates of the items' qualities, the reviewers' biases and the average variances—each against the corresponding true values. The left graphs are based on abstract values of the parameters whereas right graphs correspond to calibrated data on the Paris 1976 wine contest. We see that all points are close to the 45 degree line.

Figure 1: Monte Carlo simulations: Our Estimates versus True Values.



Notes: All graphs are generated with $n = 1{,}000$, $m = 100$ and $f = .5$. Left graphs use: $\underline{q} = 0, \overline{q} = 100, \sigma_b = 10, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. Right graphs use parameter values calibrated from the Paris 1976 Cabernet wine contest (see below Section 3.4): $\underline{q} = 9.86, \overline{q} = 14.02, \sigma_b = 1.175, \underline{\sigma} = 1.67$, and $\overline{\sigma} = 4.17$.

Next, we measure the efficiency of our estimates and investigate how they compare to

[18] $n = 10$, $m = 11$, $f = 1$, $\underline{q} = 9.86, \overline{q} = 14.02, \sigma_b = 1.175, \underline{\sigma} = 1.67$, and $\overline{\sigma} = 4.17$.

the average rating. The comparison measure—which we refer to as *Gain*—is defined as the extra share of the per item quality that is explained by our method compared to the average ratings

$$\text{Gain} = 1 - \frac{\sum_i \frac{\left(q_i^{two} - q_i\right)^2}{n}}{\sum_i \frac{\left(q_i^{avg} - q_i\right)^2}{n}}. \tag{14}$$

When there is no bias, the average ratings also converge to the true values, so this provides a ratio of how much improved efficiency is obtained by our weightings.

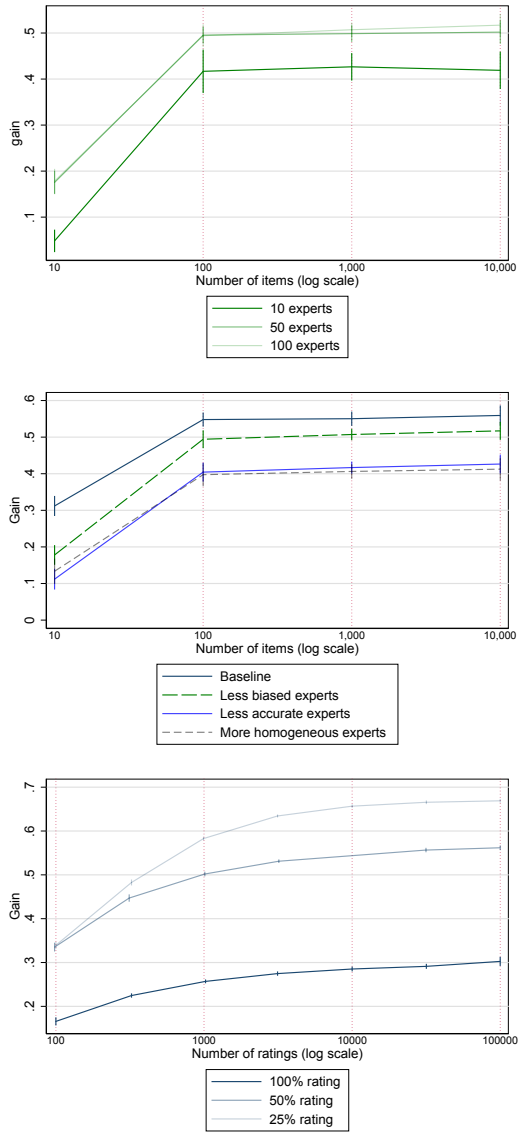
Figure 2 presents this measure for a series of Monte Carlo simulations, varying parameters. In the top graph, we vary both the number of items and the number of reviewers. We see that gain increases in both dimensions. This reflects two forces: more items enables us to learn more about each reviewer and better estimate their bias and accuracy, and so better estimate qualities, while more reviewers gives us better estimates of the qualities just through having more draws of scores. The gain relative to average scores is substantial with only 50 reviewers and 100 items, approaching 50 percent. Above that, increasing either the number of items or the number of reviewers leads to a stabilizing gain over the average rating.

In the middle graph, we keep the number of reviewers fixed and vary biases and accuracies across different numbers of items. The main insight is that reviewer accuracy is important in resolving noise in the data. It is not necessary that all reviewers be accurate, which is why increasing the homogeneity in reviewers' accuracies lowers the gains even more than does decreasing their accuracy on average. The model also performs better compared to the average when reviewers are more biased, reflecting the fact that our approach allows us to estimate and adjust for those biases whereas the average just relies on them averaging out.

In the bottom graph, we compare different random assignments of which reviewers rate which items with the same number of edges (rating observations). When the ratings matrix is sparser (reviewers rate only a limited number of items), the gain of our estimation over the average rating is larger (up to about 60% with only 1,000 ratings and up to nearly 70% with larger data sets).

We also compare our estimates to using a Borda score. The Borda score is not intended to provide an estimate of quality (an index), but a ranking. We thus transform real quality and estimated quality into ranks, and then compare them to the Borda rankings. The standard measure of rank correlation is the Spearman rank correlation coefficient. Let ρ^{two}, ρ^{B} and ρ^{avg} denote the Spearman rank correlations of our estimated quality, the Borda score and the average rating with the true quality, respectively. Table 2 shows how often each measurement offers a strictly more correlated ranking with the true ranking than each other measurement. With just 10 items, our estimates offer better matches to the true ranking than both the Borda score and the average rating, and Borda beats out the average ranking. As we increase the number of items, our estimates substantially outperform the other methods. It is already better than the two other rankings in 88% and 95% runs respectively with 100 items, and

Figure 2: Monte Carlo numerical experiments.



Notes: Fractional polynomial estimates and 95% confidence intervals where each graph point corresponds to 500 Monte Carlo data simulations. In the top graph, $f = 0.5$, $\underline{q} = 0, \overline{q} = 100, \sigma_b = 10, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. In the middle graph, the baseline corresponds to $f = .5$, $m = 100$, $\underline{q} = 0, \overline{q} = 100, \sigma_b = 20, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. The other series differ from the baseline in the dimensions specified only. The "Less biased reviewers" uses $\sigma_b = 10$. In the "Less accurate reviewers" case, we assume $\underline{\sigma} = 10$ and $\overline{\sigma} = 30$, whereas in the "More homogeneous reviewers" case, $\underline{\sigma} = 10$, and $\overline{\sigma} = 20$. In the bottom graph, we keep $n = m$, and let $\underline{q} = 0, \overline{q} = 100, \sigma_b = 10, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. For each data point, the number of experts and the number of items are adjusted to match the corresponding number of rating observations according to the horizontal axis. The percentages indicated in the legends correspond to the proportion f of the ratings matrix that are filled.

then is *always* better with 1,000 items and above. The Borda score does as good as the average rating with 10 items but performs increasingly better with more items up to 99% with 1,000 items.

Table 2: Comparing methods based on how often the given method provides a better ranking with true quality than an alternative method

# Items	$\rho^{two} > \rho^{B}$	$\rho^{two} < \rho^{B}$	$\rho^{avg} > \rho_{B}$	$\rho^{avg} < \rho_{B}$	$\rho^{two} > \rho^{avg}$	$\rho^{two} < \rho^{avg}$
10	.53	.42	.47	.49	.50	.40
100	.90	.10	.14	.86	.95	.05
1,000	1	0	.01	.99	1	0

Notes: All parameters but the number of items are calibrated from the Paris 1976 Cabernet wine contest: $m = 11$, $f = 1$, $\underline{q} = 9.86, \overline{q} = 14.02, \sigma_b = 1.175, \underline{\sigma} = 1.67$, and $\overline{\sigma} = 4.17$. Frequencies of Monte Carlo simulations for which the condition at the top of each column is respected. Frequencies for ties can be easily deduced. Calculated for 1,000 Monte Carlo simulations for each design.

5 Wine Experts' Ratings of "en Primeur" Bordeaux Wines: Estimating Wine Qualities and Experts' Biases and Accuracies

Fine wines, and Bordeaux wines in particular, have attracted much interest from economists who aim to identify wine quality and its determinants (Ashenfelter, Ashmore and Lalonde, 1995; Ashenfelter, 2008; Dubois and Nauges, 2010; Friberg and Gronqvist, 2012; Hilger, Rafert and Villas-Boas, 2011). Wine is a typical product for which quality differences are simultaneously large (e.g., prices vary significantly) and variable even for the same producer (e.g., particular wine prices vary significantly from year to year, and even within year for different wines released by the same producer, and there are many producers). The high variability of ratings by experts within and across items (Ashton, 2012; Hodgson and Cao, 2014; Ashenfelter, Ashmore and Lalonde, 1995) makes this an application in which our method of deducing quality from a set of ratings particularly valuable. In this section, we use a new dataset of ratings of Bordeaux Wines by wine tasting experts. Key parts of the Bordeaux fine wine industry operate via a futures/forwards market. At specific points in the season, wines that are not yet bottled are tasted and rated by trained professionals and experts. Their ratings are vital for intermediaries and investors who buy most of the production. Many of these ratings are eventually published in various media (magazines, books, websites). The wine is bottled and transferred to the buyers one to several years later (depending on the aging policy of the producer). Our empirical study focuses on such ratings of "en primeur" wines because these ratings are less likely to be polluted by cross influences

and other information, as they are the first ratings and are essentially simultaneous: in a given week in April critics taste the wines produced from the harvests that occurred in September and October of the previous calendar year, and their evaluations are published late April and May.

5.1 Data

Our database contains 52,968 "en primeur" ratings from 19 experts. They are wine critics, journalists, writers, and bloggers. Some like Robert Parker and Jancis Robinson are world-renowned critics. After some cleaning of the data we end up with 51,363 ratings.[19] We then delete 5,917 wine/vintages that are rated by only one expert. We end up with 45,446 ratings of $n = 6{,}346$ wine/vintages (with vintages from 1994 to 2015) given by the $m = 19$ experts. Figure D.2 in Online Appendix D presents the distribution of wines and ratings across vintage years.

5.1.1 Scaling the Ratings

The experts' wine ratings data have numerous nice properties but come with the added issue that different wine experts use different scales for their ratings. For instance, Parker rates wines from 50 to 100, but essentially only ever rates between 70 and 100. Jancis Robinson employs a scale from 1 to 20 and usually rates between 10 to 20. To adjust for these different scales we first convert all experts ratings to lie on a 0 to 100 scale and to use the whole scale. We then linearly rescale each expert's ratings so that their lowest rated wine is given a rating of 0 and the highest rated wine is given a rating of 100.[20]

Letting G denote the raw scores of the experts, the rescaled ratings are:

$$g_{ij} = 100 \times \left(G_{ij} - G_j^L\right) / \left(G_j^H - G_j^L\right), \tag{15}$$

where G_j^L and G_j^H denote j's respective lowest and highest percentiles raw rates that are used.

Figures C.1 and C.2 in Appendix C plot the distribution of ratings by experts. Given that some experts use a coarser scale than others, there are obvious peaks in their distribution. For instance, if they use a 20 point scale with half points rather than 100 point scale, then 19.5 becomes 97.5, 19 becomes 95, etc., and so there are clumps at certain points on the 100 point scale that we use.[21]

[19] Some ratings are duplicates—the same rating of the same wine and vintage by a given expert—in which case one randomly chosen is kept. Sometimes the reviewer provides intervals rather than a unique number, and then we use the mean value. The analysis is also robust to dropping the bottom five percent of the wines.

[20] Confining the wines to a finite scale takes us outside of the model, but given the large number of ratings, the extreme ratings are far in the tails of the distributions, and so there is no censoring of data.

[21] See Section 7 for a brief discussion about grids.

The lower tail of ratings is long and noisy, and so we have run the same analysis when the lowest five percent of ratings are dropped. Unreported results show it makes little difference given the size of this dataset. However, this could matter in settings where there are rare but peculiar outliers that then distort the scale. More generally, if there is selection in which items a reviewer rates, then one has to adjust for that in the normalizations. For instance, if most reviewers rate all items, but some particular reviewer only rates high-quality items, then rescaling that reviewer's scale to match the others would affect that reviewer's estimated bias and variance. We do not have to address this issue in our application since there are many Bordeaux wines that are prominent wines and all the experts 'have' to rate, and they cover a full spectrum of ratings, and so all of the reviewers are rating a fairly full spectrum of quality. We show those distributions in Online Appendix D, Figure D.1. Only one reviewer, Jacques Perrin, might be suspected to have left censured ratings with respect to quality. As we have a limited number of reviews from Jacques Perrin (488), his normalization has little impact on the overall analysis.[22]

5.2 Estimating Experts' Biases and Accuracies

In Table 3 we summarize experts' estimated characteristics. Figures D.3 and D.4 in Online Appendix D present that same information in more detail.

As the accuracy $\left(\sigma_j^{two}\right)^{-2}$ is hard to interpret directly, we normalize by multiplying it by the average variance of the experts, $\sum_{j'}\left(\sigma_{j'}^{two}\right)^2/m$. The estimated normalized accuracy of expert j is noted A_j^{two}. Thus, an expert with the average accuracy would show up as having accuracy 1. An expert with accuracy 2 has twice the average precision, and so forth.

We can also measure how correlated an expert's ratings are with the estimated true quality of the wines s/he rates. The correlation of an expert's prediction of the quality of a wine is related to the expert's accuracy, as we now describe. Let σ_q^2 be the variance in the quality of a typical wine. Note that

$$Cov(q_i, g_{ij}) = Cov(q_i, q_i + b_j + \varepsilon_{ij}) = Var(q_i) + Cov(q_i, \varepsilon_{ij}) = \sigma_q^2.$$

Therefore,

$$Corr(q_i, g_{ij}) = \frac{Cov(q_i, q_i + b_j + \varepsilon_{ij})}{\sigma_q\sqrt{Var(q_i + b_j + \varepsilon_{ij})}} = \frac{\sigma_q^2}{\sigma_q\sqrt{\sigma_q^2 + \sigma_j^2}} = \left(1 + \frac{\sigma_j^2}{\sigma_q^2}\right)^{-\frac{1}{2}}.$$

[22]This could be an issue in other data sets in which many reviewers choose to rate only parts of the distribution. An approach to dealing with this is to begin as we do here and develop estimated qualities on all of the items. Then, based on those qualities, one can redo the normalizations if some reviewers are only estimating part of the quality distribution, to account for that (for example if some reviewer only rates items that have estimated quality above 50, then one would normalize their ratings to the interval 50 to 100, and so forth).

Thus, since accuracy is $\frac{1}{\sigma_j^2}$ and correlation is $\left(1+\frac{\sigma_j^2}{\sigma_q^2}\right)^{-\frac{1}{2}}$, the two are similar functions.[23] We study the relationship between accuracy and correlation and find a positive relation between the two indicators, but they are clearly distinct (see Figure D.5 in Online Appendix D).

Recall that our model presumes that the experts' accuracies are independent of the quality of a wine - so they are just as accurate at rating a high quality wine as a low quality wine. In essence we assume that $q_i \perp \varepsilon_{ij}, \forall i, j$. One might expect that experts' errors would increase when wines are of lower quality; or one might even expect the opposite. We study the relation between the estimated wine qualities and errors in Online Appendix D (Figure D.8). We see little relationship between errors and quality from the fifth percentile of item quality.

Table 3: Experts' Accuracies and biases.

Expert	$\left(\sigma_j^{two}\right)^2$	A_j^{two}	Corr (g_{ij}, q_i^{two})	b_j^{two}	n_j
Antonio Galloni	72.41	0.96	0.79	1.45	1,140
Bettane et Desseauve	64.46	1.14	0.82	7.98	3,011
Chris Kissack	88.19	0.76	0.79	0.90	2,431
Decanter	65.11	1.07	0.88	-9.83	2,342
Jacques Dupont	149.77	0.41	0.69	-18.32	3,077
Jacques Perrin	91.71	0.70	0.89	-22.13	488
James Suckling	81.56	0.83	0.82	-3.30	1,985
Jancis Robinson	80.59	0.86	0.69	-1.26	3,793
Jean-Marc Quarin	39.62	2.21	0.87	0.57	3,042
Jeannie Cho Lee	51.00	1.53	0.84	14.42	1,308
Jeff Leve	67.32	1.03	0.89	-3.07	1,530
La Revue du Vin de France	91.72	0.72	0.80	-1.63	2,216
Neal Martin	52.22	1.52	0.82	12.01	2,965
Rene Gabriel	70.82	1.03	0.80	8.96	4,757
Robert Parker	56.78	1.41	0.81	13.14	2,838
Tim Atkin	73.69	0.97	0.75	7.51	1,900
Wine Enthusiast	118.89	0.52	0.75	-5.10	2,513
Wine Spectator	74.84	0.92	0.84	5.45	3,669
Yves Beck	141.53	0.44	0.78	-7.74	441

5.3 Estimating Wine Qualities

We present the top-100 wines from the sample along with their estimated qualities in Table C.1 in Appendix C. The number one Bordeaux wine is actually a Sauterne (sweet white

[23]Note that this correlation is not estimable without using our method, since one needs to estimate the quality of the wines to estimate the correlation of an expert's ratings with that quality.

wine), Chateau Yquem 2009, and Chateau Margaux 2010 is the best red wine.[24]

As our qualities use the full 100 point scale and have an average in the 30's, the reported qualities may "look" unfair as most of the consumers and experts have the most known experts' ratings distribution in mind. For instance, most people have an idea of what an 80 or 90 point rating of a wine means according to Robert Parker. For instance, it might seem strange to any professional in the fine wine industry to give a less than 90 point rating to a Lafite Rotschild 2010. To avoid potential misunderstanding due to interpreting wine qualities in the scales that people are often used to, we also rescale our quality ratings to place them back in the subregion of the 100 point scale usually used by wine experts – who rate almost all wines between 70 and 100. To do this, we also calculate a "Parker-equivalent" quality level that uses the same part of the scale that Parker usually uses. Figure D.6 in the Online Appendix D shows how the distribution of ratings on the 100 points scale is modified when rescaled to a "Parker nominal view". Note that this of course does not modify at all the ranking of the wines - it is just a shifting and renormalizing of the scale. This modified quality is reported in the second column (entitled "rescaled") of Table C.1 in Online Appendix C.

5.3.1 Monte Carlo Simulations Calibrated to Bordeaux Wine Data

The Monte Carlo simulations above show that the methodology performs well in resolving noise in the rating, across various abstract circumstances. We now run Monte Carlo simulations calibrated to the larger set of Bordeaux wine data. This involves extracting parameter information on experts biases and their accuracy as we have done in Section 3.4 on the 1976 Paris contest. Unlike in the 1976 Paris contest, experts do not rate all wines but only some of them. Therefore, we use exactly the rating structure stemming from the data: not only the same number of items and the same number of experts, but also the same mapping between those two sets (the "who rates which item"). Detailed results are presented in Online Appendix D.2 (Table D.3 and Figure D.7). The average fitness is 86% and the average gain with respect to the mean rating is 41%, which is consistent with what has been found in the other Monte Carlo simulations.

5.4 Red wines

As Bordeaux wineries are best-known for their red wines, we also report a separate ranking restricted to that subsample. The results are presented in the Online Appendix, see Tables D.4 and D.5 and Figure D.9.

[24] Once again, we emphasize that the discussion from Section 2.3 apply here about interpreting our "quality" estimate as an anchor that best predicts an infinite population's *subjective* ratings.

5.5 Biases and Accuracies that Vary with Categories of Items

Any reviewer's ability and judgment in rating items might vary with categories of items. There is no reason to expect that an expert who is extremely accurate in reviewing wines would be a good analyst for recommending movies or cars or stocks. Where do such distinctions end? It might be that an expert on wines is much better at judging red wines than white wines, or judging Bordeaux wines than Spanish wines. The distinctions do not end there: even within Bordeaux there are distinctly different red wines. The wines from the "left bank" (the west side of the Gironde Estuary) and the "right bank" (the east side), generally contain different blends of grapes and come from different soils and can even have different weather conditions. The left bank wines are blends that predominately feature Cabernet Sauvignon grapes, while the right bank wines tend to feature Merlot grapes, with varying mixtures and often including Cabernet Franc and other grapes. While not as different as red from white, there are still sufficient distinctions that make these two categories different from each other and it can be that a given expert would favor Cabernet Sauvignon over Merlot grapes, or vice versa. This might result in different biases and/or accuracies for the two regions.

Effectively any given expert can be treated as two completely different experts, one for Left Bank Bordeaux and one for Right Bank Bordeaux.[25] One of those two experts might have a large positive bias and the other a slight negative bias, and correspondingly one might be very accurate and the other more variable. One could interpret the biases as "preferences" expressing a particular taste of the reviewer: a deviation from the average "true" quality that favors or goes against a certain type of wine. Thus, for any given set of items N, we can partition that set, and treat every distinct group as a completely different set of items and run our algorithm separately on that set of items. Thus, for every reviewer we end up with a different bias and accuracy for every category of items.

To illustrate this, and to see that Left Bank and Right Bank wines are actually quite distinct in terms of experts' biases and accuracies - we do this by dividing our data on Bordeaux wines.

Left vs Right Bank Tastes of Experts Let L denote "Left Bank" and $N\backslash L$ denote "Right Bank", and let k generically refer to observable product categories. Formally, the evaluations of any expert j are now category-dependent:

$$g_{ij} = q_i + b_{j,k} + \varepsilon_{i,j,k}, \quad \forall i \in k, \forall k \in \{L, N\backslash L\} \tag{16}$$

Thus, experts have category-specific biases that are interesting to compare. The differ-

[25] As an extension, one could instead assume that the experts have similar accuracies but different biases across the regions. In our setting, we have many ratings by each expert and so can get precise estimates of their biases and accuracies from each region separately, but for settings with less information presuming some relationship of biases and accuracies across product categories could improve the power of the estimation.

ences in the estimated biases across the left vs right dichotomy are:

$$\Delta b_j^{two} = b_{j,L}^{two} - b_{j,N\setminus L}^{two}, \tag{17}$$

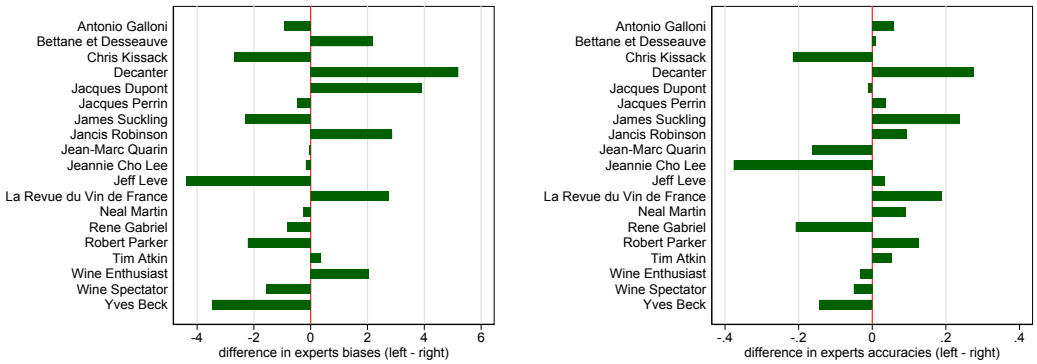
and similarly differences in expert j's normalized accuracies between Left and Right Bank wines as

$$\Delta A_j^{two} = A_{j,L}^{two} - A_{j,N\setminus L}^{two}, \tag{18}$$

where $A_{j,k}^{two}$ is the normalized estimated accuracy of expert j for category k.

These are pictured in Figure 3.

Figure 3: The biases and accuracies of experts when their specific biases for left bank (vs right bank) wines are taken into account.



We can see that Robert Parker is a "rightist," which is consistent with him being known for advocating in favor of powerful Bordeaux wines, mostly located on the right bank. Other pronounced "rightists" include Jeff Leve, James Suckling, Chris Kissack, Wine Spectator and Yves Beck. On the other side, Decanter, Jacques Dupont, La Revue du Vin de France, Jancis Robinson, Wine Enthusiast, and Bettane et Desseauve are sometimes said to favor more traditional and reserved wines. Some bias positively Left Bank wines, such as Jancis Robinson (also Decanter, Jacques Dupont, La Revue du Vin de France and Wine Enthusiast). This could explain partly the well known Pavie 2003 controversy[26] and more generally the lack of correlation between Parker's and Robinson's ratings which is presumed to be due to different preferences in wine "styles" (Ashton, 2016). Some experts do not bias one category of items over the other.

Interestingly, there is no clear correlation pattern between the differences in accuracies and the differences in biases (and see Figure D.10 in Online Appendix D for more detail). It is not because an expert gives a "premium" to a given type of red wine that this expert is found to be more or less accurate in rating those wines.

[26]See https://www.sfgate.com/wine/article/Robinson-Parker-have-a-row-over-Bordeaux-2755642.php.

A Significant Difference We can test whether there is a significant difference in Left Bank and Right Bank wines by examining whether there is a significant improvement in the predictions of qualities when distinguishing wines from the two areas.

First we define the residual weighted sum of squares for the different ways of estimating. Without any distinction between Left and Right Bank wines, the overall weighted sum of squared errors from keeping all the wines in one category was:

$$RSS_1 = \sum_{i,j} 1_{ij} \left(g_{ij} - b_j^{two} - q_i^{two} \right)^2 A_j^{two}. \tag{19}$$

The adjustment by A_j^{two} weights the terms so that the errors are normalized to have the average variance and thus the same distribution - which is the same as weighting each estimate by its relative precision which produces the overall estimated sum of squared errors. Since

$$\sum_{i,j} 1_{ij} \left(g_{ij} - b_j^{two} - q_i^{two} \right)^2 / \left(\sigma_j^{two} \right) = R$$

this becomes

$$RSS_1 = \frac{R}{m} \sum_{j'} \left(\sigma_{j'}^{two} \right)^2 \tag{20}$$

Once we divide things into two categories, we end up with a second sum of squared errors:

$$RSS_2 = \sum_{i \in L,j} 1_{ij} \left(g_{ij} - b_j^{two} - q_i^{two} \right)^2 A_{j,L}^{two} + \sum_{i \in N \setminus L,j} 1_{ij} \left(g_{ij} - b_j^{two} - q_i^{two} \right)^2 A_{j,N \setminus L}^{two}.$$

Using the similar calculations as for Equation 20, it comes:

$$RSS_2 = \frac{R_L}{m} \sum_{j'} \left(\sigma_{j',L}^{two} \right)^2 + \frac{R_{N \setminus L}}{m} \sum_{j'} \left(\sigma_{j',N \setminus L}^{two} \right)^2, \tag{21}$$

with R_L ($R_{N \setminus L}$) the number of ratings of Left Bank (Right Bank) wines and noting that all experts are rating wines on both Left and Right Banks, and so there is no subscripting on m.

We end up with $37,982$ ratings of red wines for which the Left or Right bank is clearly identified (some wines blend grapes from both sides of the river and the origins of some others is not clear in the data). These divide into $n_L = 20,266$ ratings of Left Bank wines and $n_{N \setminus L} = 17,716$ of Right Bank wines. Then, with our data, we find $RSS_1 = \frac{37,982}{19} \times 1,529.154 = 3,056,860$, and $RSS_2 = \frac{20,266}{19} \times 1,415.752 + \frac{17,716}{19} \times 1,593.419 = 2,995,823$.

There are 38 parameters estimated in the original algorithm and 76 parameters estimated in the algorithm in which we split wines into Left and Right Banks. This results in an F-test statistic of:

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{76-38} \right)}{\left(\frac{RSS_2}{36,821-76-1} \right)} = \frac{\left(\frac{61,037}{38} \right)}{\left(\frac{2,995,823}{36,744} \right)} = 20.24$$

At a 1 percent significance level, the F-test threshold with $(38; 36,744)$ degrees of freedom

is 1.59. We see that our F statistic of 20.24 greatly exceeds that threshold value. Thus, there are significant differences in experts' rating patterns for Left and Right Bank wines.[27]

6 Bordeaux Wines' Estimated Qualities, Experts' Accuracies and Prices

In this section, we examine the relationships between our estimated wine quality and prices. If one believes that markets learn items' qualities,then our wine quality estimates should be positively correlated to prices controlling for a number of other confounding factors.

The view point can be reversed to take an IO perspective. If one believes our technique really captures item quality, and thus experts' accuracies and biases, then this section contains results on demand reactions to quality variations.

We observe posted prices of the rated wines in retail shops in three major markets across the world. Generally, there is a textbook identification problem (e.g., see Working, 1927) that stems from the fact that prices are determined by both supply and demand, which can both move to affect prices. Here, identification comes from the fact that prices are largely determined after the amount of each wine supplied is already largely fixed, and then the quality of the wine is later made known and prices result. Thus, we treat supply as inelastic, and prices reflecting perceived quality. Moreover, by including various fixed effects, it is deviations in prices that are being attributed to relative qualities of the wines.

6.1 Prices and Other Data

Data on Wines, Official Rankings and Vineyards The Bordeaux wine "terroir" is typically documented by sub appellations such as Medoc, Saint Emilion, Premieres Cotes de Bordeaux or Pauillac. These appellations relate to specific sub-regions of production as well as some production constraints (types of grapes, upper bounds on production quantities per hectares of vineyard, selection of vineyards...). Our dataset contains this information for each wine (see Table D.1 in Online Appendix D). We also know when wines are "first wines" (their top wine, if they make more than one) of a "chateau" that was listed in one of the official rankings of the Bordeaux production area, such as Grand Cru Classé 1855 or Premier Grand Cru Classé A (see Table D.2 in Online Appendix D). We can use these data as controls.

Prices and Store/Market Data The prices of the wines are from surveys of restaurants in three of the main world-wide markets: Hong Kong, New York and Paris. In these cities, the

[27]More generally, introducing added categories and allowing reviewers' biases and accuracies to vary by those categories is a classic specification problem. Adding dimensions risks over-fitting, and standard techniques that penalize the addition of new dimensions can be used to assess whether dimensions should be added. We do not pursue that question here.

wine prices of, respectively, 244, 437 and 409 restaurants were surveyed at different points in time (details appear in Table E.1 and Online Appendix E). The prices were recorded between 2010 and 2016. Initially, 93,466 prices of standard bottle Bordeaux wines were recorded.

The Data Merge We match each wine/vintage rated en primeur with all posterior prices and obtain a database of wine/vintage-price observations in a given shop and year. Out of the 2,871 wine/vintages that we consider, we have 43,307 such observations, that is 15.08 prices on average for each wine/vintage.

In Online Appendix E, Figure E.1 shows the price distributions in the three markets and Table E.2 lists the top-20 most surveyed restaurants in the data.

6.2 Do Estimated Qualities Predict Prices?

In the Bordeaux wine industry (as for other AOC in France), quantities are largely fixed for any given vintage.[28] The main adjustment to an increased individualized demand is thus on the price and we therefore estimate an hedonic (price) regression to appreciate whether and how estimated quality affects the demand of given wines.

We cannot however simply regress prices on our estimated quality because other factors influence the posted prices. For instance, shops' attributes, vintages, local production origins and official rankings can be observed by consumers—who may have various levels of information about quality—and so may affect the prices, holding wine quality constant. We thus include appellation and official ranking fixed effects as well as retail shop fixed effects which can influence the observed prices. Sale year dummies are also considered as it captures yearly wine market and more global economic conditions.

In addition, Ashenfelter, Ashmore and Lalonde (1995) and Ashenfelter (2008) highlight that wine yields and prices are affected by weather conditions at crucial points in the season in the production year. We also control for such weather conditions by including vintage-appellation fixed effects: dummies that capture the weather conditions for various vintages in the specific sub-regions of Bordeaux production. Given that weather can also be highly correlated with wine quality, we expect that this will lead to an underestimation of the effect of wine quality on prices.

More importantly for our purpose, consumers may also be directly influenced by some experts' ratings.[29] Omitting such variables could lead prices to correlate with our estimated quality simply because our quality estimates are also positively correlated with key expert

[28] Production cannot be significantly adjusted upward by mixing the wine of a vintage with wine from other vintages since at least 85% of the wine must come from property grapes of the referenced vintage. There are occasional weather disasters that lower quantities, but the quantities supplied are effectively inelastic in the short run.

[29] Information salience has been discussed in the context of taxation by Chetty, Looney and Kroft (2009), of college rankings by Luca and Smith (2013) and of consumers online rating by Luca (2016). In the wine industry, Ali, Lecocq and Visser (2008) used a natural experiment to show that Parker ratings have a direct and significant impact on prices.

ratings that consumers and wine shops managers observe. In essence, our problem reverses a traditional question addressed in the wine economics literature which aims to identify the impact of the ratings on the prices when wine quality is unobserved by the econometrician. Instead, we estimate the relationship between wine quality (estimated by our technology) and prices controlling for salient information.

In particular, Ali, Lecocq and Visser (2008), Dubois and Nauges (2010), Friberg and Gronqvist (2012), and Hilger, Rafert and Villas-Boas (2011) have found that well-known experts ratings have a direct impact on prices (while controlling for quality using different empirical strategies). We therefore control for the salient experts' ratings by directly including the ratings of the best-known expert for Bordeaux fine wines, Robert Parker. We also include the ratings of Jancis Robinson, who is another big name for Bordeaux wines. In some regressions, we also control for the "best" rating of each wine as in retail stores, sellers often transmit to consumers the most favorable rating(s) so as to influence consumers' decisions, and may thus take this information into account in the pricing.[30] Lastly, we allow for variation in ratings to affect the perception of wine quality by controlling for the variance in the ratings of each wine.[31]

Section G.1 in the Online Appendix proposes possible micro-foundations for the price reactions to quality. In the spirit of Card and DellaVigna (2017), we model consumers (it could be the restaurant sommelier or the retail wine manager) receiving a noisy signal of wine quality, and observing fundamentals (official ranking, appellation, ...) as well as the rating of some reference expert.

Results (see Table 4) show that our estimated quality is a strong predictor of prices as its coefficient is large, positive and significant at the .001 level in all regressions. Note that a number of fixed effects have been included such as vintage×appellation, official ranking, price year, variance of ratings of each wine, and store fixed effects. Regressions in which we do not introduce all those controls (see Online Appendix Table E.3) show that introducing estimated quality in the regression raises the R^2 from .20 to .60. In particularly telling regressions (Columns 4–7), in which the best rating or ratings of famous experts, Robert Parker and Jancis Robinson, are included as regressors, our estimated quality significantly predicts the price. Interestingly, the other salient ratings (except Robinson's) end up having little significant influence on prices after controlling for our estimated quality, even though they have been found to significantly affect prices in previous studies (that do not include our quality estimates).

As prices and ratings are in logs, the coefficients In the first column of the table our index has a large positive coefficient of 3.2, which is consistent with the idea that, in the fine wine market, prices are highly sensitive to quality variations—a 10 percent increase in quality raises the price by nearly 32 percent.

[30] All ratings used are normalized to span the 0-100 scale (as discussed by Equation 15).

[31] We cannot use the average rating among experts as a supplementary control as that fails VIF multicollinearity tests (whereas other regressions pass this test).

In the last two columns of Table 4, we compare how our estimated quality and average rating predict prices depending on the time over which prices are collected. In Column 8, where prices formed up to ten years after harvesting, both our estimated quality and the average rating are comparable and significant. By contrast, in Column 9 where we only consider the prices of wines that are less than five years old (rated in the last four years) we find that our estimated quality has more of an influence on those prices whereas the average rating has no significant effect in the first five years. Thus, the average rating only emerges later; which would be consistent with the dissemination of those ratings affecting the remaining demand for the wines over more time.

Table 4: Retail prices as a function of estimated wine quality and of salient and best en primeur ratings.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Estimated quality	3.242^{+}		2.451^{+}	2.411^{+}	2.534^{+}	2.717^{+}	2.804^{+}	1.415^{+}	2.149^{+}
	(21.18)		(15.70)	(10.48)	(10.92)	(16.51)	(11.47)	(4.09)	(6.02)
Best rating				0.0569			-0.194	-0.139	0.142
				(0.29)			(-1.10)	(-0.79)	(0.78)
R. Parker rating					0.0626				
					(0.28)				
J. Robinson rating						$0.189^{\#}$			
						(2.58)			
Average rating								1.382^{+}	0.275
								(4.68)	(0.99)
N	43231	43215	43215	43215	36017	28673	22149	22149	4921
r2	0.681	0.722	0.802	0.802	0.798	0.822	0.796	0.798	0.809
aic	76017.4	70096.6	55488.9	55488.7	45748.6	31686.1	24113.6	23908.3	4057.7
bic	76034.7	70105.2	55514.9	55523.4	45782.6	31719.1	24145.6	23948.3	4090.2

Notes: t-statistics are in parentheses. The standard errors are two-way clustered at the wine×vintage level and at the store level. Significance levels: $^{\#}p < 0.1$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All ratings are corrected to span a 0-100 scale (see Equation 15). The price variable and the listed variables are all in logs so that coefficients can be interpreted as elasticities. All regressions but Column 2 include the variance of the considered wine ratings. All columns but Column 1 include vintage×appellation and official ranking fixed effects. All models include a number of fixed effects: year, type (red, white or sweet), market and store fixed effects. Columns 7 and 8 consider only the prices of wines formed less than 10 years after production (thus 9 years after the "en primeur" reviews), and Column 9, only the prices of wines formed less than 5 years after production.

6.3 Are the Ratings of More Accurate Experts Better Predictors of Prices?

We have shown that estimated wine qualities are correlated with retail prices, controlling for many things (including salient ratings). This tends to confirm that prices do reflect

Bordeaux wine quality. It is also providing external support to our item quality estimation methodology. We now assess another important output of our methodology: the estimation of reviewers' accuracies. Are estimated expert accuracies also consistent with price data? This is important because consistently estimating reviewers' accuracies is key to our method, and this provides external validity that those estimates are predictive.

We expect more accurate experts to have greater correlation of their ratings with prices because their ratings capture more strongly item quality, which in turn likely correlates positively with prices –as we have seen in the previous subsection. There are however other factors which may affect prices, besides wine quality, which may also be correlated to wine quality. To isolate the correlation between prices and experts' ratings from external confounding factors, we first regress log prices on each expert's log ratings separately, controlling for a number of dummy variables such as the rating year, the year, the interaction of vintage and appellation dummies (which captures in particular local weather conditions in the production year), official ranking, wine type (red, white or sweet), market, and retail shops fixed effects. Raw regression results appear in Table E.4, Online Appendix E). As both ratings and prices are in logs, we can interpret those coefficients as each expert's ratings-elasticity of wine prices. Of course, this is not a causal relationship, as it also reflects, for instance, that more accurate experts are more correlated with quality which correlates with price.

In a second analysis, we study the relation between those estimated elasticities and experts' accuracies. Among the thirteen experts considered,[32] the most accurate expert, Jean-Marc Quarin is also the one whose ratings correlate most with the prices. A 10 percent increase in his ratings corresponds to a 25.4 percent increase in prices. Robert Parker, who is the second most accurate in this list, has the second highest correlation between ratings and prices: a 10 percent increase in his ratings corresponds to a 19.5 percent increase in prices.

Figure 4 plots experts' estimated accuracies against their ratings-elasticity of prices.[33] We see that there is a clear positive relation between the two. Most experts lie within the 95% confidence interval of a linear prediction with a (nearly) unitary slope. This is consistent with the idea that the correlation between an expert's ratings and prices increases with estimated expert's accuracy. Some experts lie above or below the confidence interval. Bettane et Dessauve and Robert Parker have a correlation with prices that goes beyond what is predicted by their accuracy. Some others—Neal Martin, Tom Atkin and Decanter—have a correlation with prices below what is predicted by their accuracy. This *residual* correlation could reflect different things. Here are two possibilities. It could be that the expert's rating influences the price, as is often claimed, for instance, about Parker's ratings.[34] It could also

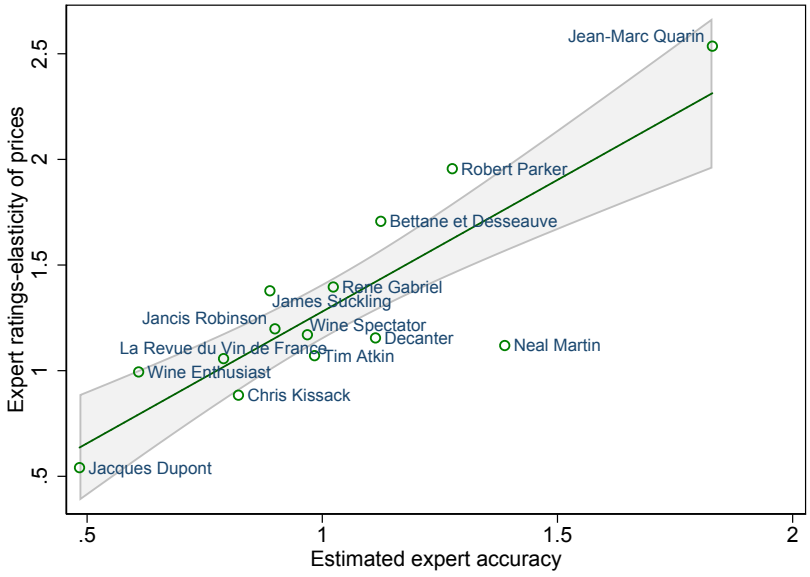
[32]Six experts (Antonio Galloni, Jacques Perrin, James Suckling, Jeannie Cho Lee, Jeff Leve, Yves Beck) could not be considered as too few of their ratings were for wines with observed prices (regressions do not converge given the large number of fixed effects introduced).

[33]We do not control for our quality estimates because those are formed from the experts' ratings, and so someone who was a perfect match of quality would then have no impact on prices.

[34]For example, see "Do Wine Scores Matter? James Suckling's retirement from Wine Spectator will tell

be that the expert's rating is affected by the anticipated price point that a wine will sell at – giving higher ratings to more expensive wines (after adjusting for quality).

Figure 4: Experts' accuracies against the coefficients of their ratings regressing wine prices.



Notes: The vertical axis lists the estimated coefficients of each expert's ratings, regressing retail prices on each expert's ratings, controling for the rating year, vintage×appellation, official ranking, wine color, and retail shop fixed effects. In those regressions, prices and ratings are in logs, so that coefficients can be interpreted as elasticities. Regressions did not converge for Jacques Perrin and Yves Beck. Some did converge, but the focal expert rated a limited number of wines that are priced (less than 1,200) like Antonio Galloni, Jeff Leve and Jeannie Cho Lee. Their coefficients are thus not reported here. All regressions are exposed in the Online Appendix, Table E.4.

Figures E.5–E.8 in Online Appendix E show regression results obtained when prices are regressed on the ratings of most influential experts (on all markets, and on each market separately).

6.4 Re-Ratings

In Online Appendix F, we examine another aspect of our Bordeaux wine quality estimations relying on a very different design, making use of additional rating data. Some experts rate the exact same Bordeaux wines at (at least) two different points in time: a first time "en primeur" (those ratings are the ones we have considered so far to estimate wine quality), and later – usually after the wine has been bottled and is already available in retails.

us for certain" in Forbes, July 15, 2010, and *The Emperor of Wine: The Rise of Robert M. Parker, Jr., and the Reign of American Taste* by E. McCoy, Harper Collins, 2014.

We find that when experts re-rate the same wine, they correct errors in their initial ratings as if they were adjusting their previous rating to be closer to the "true" quality. Results remain when we consider that experts may also be directly influenced by other experts' opinions, and therefore control, as in the previous section, for salient rates. We also take into account the expert specifics and the other factors that influence the evolution of quality between the "en primeur" quality and the re-rating year. In addition, we show that when re-ratings come later, say more than three years after the initial rating, then the coefficient of estimated quality is larger whereas the initial rating becomes less important. However, as estimated quality is correlated with average ratings, we cannot exclude that experts are herding in some way.

7 Concluding Discussion

We have provided a technique that processes a series of ratings by a group of reviewers and simultaneously provides: unbiased and consistent estimates of the items' true qualities, together with consistent estimates of each reviewer's bias and accuracy. In applying the technique to more than forty thousand expert ratings of Bordeaux wines of vintages from 1998 to 2015, we obtained estimates of prominent experts' biases and accuracies, as well as estimates of the 'true qualities' (consensus values) of the wines. We showed that our quality estimate is a significant (at the .001 level) and strong (with an elasticity of 3.4) predictor of wine prices, even when controlling for many fixed effects and other measures of ratings. The fact that our technique not only identifies estimates of item true qualities, but also provide estimates of reviewers' biases and qualities should also be valuable. For instance, one can identify the most accurate reviewers and incentivize them to provide ratings on particular items that may be in high demand. One can also weed out reviewers who are inaccurate and trying to manipulate scores. Our technique is also easily extended to allow reviewers' biases and accuracies to vary with categories of items, and so estimates can be tailored to product categories. For instance, some expert may be more accurate and less biased in rating Bordeaux than Rhone wines.

We close with notes on further extensions and applications of our techniques.

Our analysis presumes that ratings are randomly distributed around the true quality subject to reviewer bias. That is, reviewers do not deliberately lie or adjust specific reviews. However, there are instances in which reviewers have been reported to be paid or bribed to provide extreme ratings, including rating games, apps, and restaurants. In some cases, a product might even create a fake reviewer with reviews of many products to establish a history and visibility, just to review its product and provide it with an outstanding rating. Given the inherent noise in any particular review, it can be impossible to know whether any single item was deliberately biased by any single reviewer. However, there are two cases in which our technique can identify whether there are fraudulent reviews. The first is in a case in which many reviewers rate a particular item, and a nontrivial fraction but not all

of them are bribed. This case results in a pattern in which the distribution of reviews does not follow the usual random pattern around the reviewers' biased points, but instead has an extra mode at a high level with a statistically rare number of reviews that deviate from their mean. The second case is in which a given reviewer is bribed a non-trivial fraction of items. In this case, the reviewer has an abnormally high number of reviews that are outliers, given that reviewer's bias and accuracy and the true quality of the items.

Another extension concerns the fact that our model is one with continuous and uncensored scores. Many applications are ones in which experts assign discrete scores on a course and bounded grid. These are not major issues for either of our applications as the grids are fairly fine and none of the ratings are censored in the 1976 competition, and less than one percent of the ratings reach the top of ratings reach an expert's upper limit and none reach the lower limit in the larger Bordeaux ratings. Nonetheless, there are settings where people can just assign scores of 1, 2, 3, 4, 5, or something similar. These more restricted grids end up both censoring and distorting scores if true qualities are more continuous and/or have no obvious bound. An adaptation of our approach to settings in which scores are forced to finite grids is as follows. We can model this by having experts map their rating to the closest point on the grid. The probability of ties is zero, and so this is a well-defined process. This process becomes nonlinear and so one way to estimate the underlying qualities, biases, and accuracies is by simulated method of moments. In particular, for every potential profile of actual item qualities q_is and b_j, σ_j^2s, one can simulate scores by randomly drawing them according to (1) and then map them to the closest point on the grid. The simulated ratings $\widetilde{g}_{ij}$ can then be differenced from the actual ratings g_{ij}, and the combination of q_is and b_j, σ_j^2s that minimize the total sum of squared error can be found. The set of potential q_is and b_j, σ_j^2s is infinite and so has to be approximated. Even fitting them on a grid produces a large set of potential parameters. The actual average scores provide a starting estimate of the qualities, and the direct estimates of the biases and variances based on those provide starting points. These are biased due to the censoring and forcing of points onto the grids, and so it is important to search, but given the size of the potential parameter space, it is important to search in intelligent directions. One can use methods from censored regression analysis (e.g., see Tobin, 1958; Amemiya, 1973 and the literature that followed) to estimate the biases as well as the variance of the errors for any given set of qualities; but then still has to iterate on the estimation of qualities. Optimal techniques for that search process is an important issue for further research.

Finally, the precision of the estimated bias and accuracy of any given reviewer depends not only on how many items they rate, but with which other reviewers their ratings overlap. Overlapping with more other reviewers, and more accurate other reviewers, gives more precision to the estimated qualities and thus more information about any given reviewer's characteristics. This information is implicit in our estimation, but one could explicitly explore the structure of the bipartite network of reviewers and items and how that structure affects the power of the approach and precision of various estimates.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian and Asuman Ozdaglar. 2022. "Learning From Reviews: The Selection Effect and the Speed of Learning." *Econometrica* 90(6):2857–2899.

Akerlof, George. 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84(3):488–500.

Ali, Hela Hadj, Sebastien Lecocq and Michael Visser. 2008. "The Impact of Gurus: Parker Grades and En Primeur Wine Prices." *The Economic Journal* 118(529):F158–F173.
URL: *http://dx.doi.org/10.1111/j.1468-0297.2008.02147.x*

Amemiya, Takeshi. 1973. "Regression analysis when the dependent variable is truncated normal." *Econometrica: Journal of the Econometric Society* pp. 997–1016.

Arrow, Kenneth J. 1951. *Social choice and individual values*. Yale university press.

Ashenfelter, Orley. 2008. "Predicting the Quality and Prices of Bordeaux Wines." *Economic Journal* 118:F174–F184.

Ashenfelter, Orley, D Ashmore and R Lalonde. 1995. "Bordeaux wine vintage quality and the weather." *chance* 8:7–14.

Ashenfelter, Orley and Richard Quandt. 1999. "Analyzing a wine tasting statistically." *Chance* 12(3):16–20.

Ashton, R. 2016. "The Value of Expert Opinion in the Pricing of Bordeaux Wine Futures." *Journal of Wine Economics* 11:261–288.

Ashton, Robert H. 2012. "Reliability and Consensus of Experienced Wine Judges: Expertise Within and Between?" *Journal of Wine Economics* 7:70–87.

Askalidis, Georgios and Edward C Malthouse. 2016. "Understanding and overcoming biases in customer reviews." *arXiv preprint arXiv:1604.00417* .

Balinski, Michel and Rida Laraki. 2013. "How best to rank wines: Majority Judgment." *Wine Economics* pp. 149–172.

Banerjee, Abhijit V. 1992. "A simple model of herd behavior." *The Quarterly Journal of Economics* 107(3):797–817.

Bikhchandani, Sushil, David Hirshleifer and Ivo Welch. 1992. "A theory of fads, fashion, custom, and cultural change as informational cascades." *Journal of political Economy* 100(5):992–1026.

Budescu, David V. 2005. "Confidence in Aggregation of Opinions from Multiple Sources." *Information sampling and adaptive cognition* p. 327.

Budescu, David V. and Eva Chen. 2015. "Identifying expertise to extract the wisdom of crowds." *Management Science* 61(2):267–280.

Cao, Jing and Lynne Stokes. 2010. "Evaluation of wine judge performance through three characteristics: Bias, discrimination, and variation." *Journal of Wine Economics* 5(1):132–142.

Card, David and Stefano DellaVigna. 2017. "What do Editors Maximize? Evidence from Four Leading Economics Journals." *NBER Working Papers* 23282.

Cardebat, Jean-Marie, Jean-Marc Figuet and Emmanuel Paroissien. 2014. "Expert opinion and Bordeaux wine prices: An attempt to correct biases in subjective judgments." *Journal of Wine Economics* 9(3):282–303.

Carroll, Raymond J and David Ruppert. 1982. "Robust estimation in heteroscedastic linear models." *The annals of statistics* pp. 429–441.

Chetty, Raj, Adam Looney and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99:1145–1177.

Chevalier, Judith A and Dina Mayzlin. 2006. "The effect of word of mouth on sales: Online book reviews." *Journal of marketing research* 43(3):345–354.

Condorcet, M. le Marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Reprinted, Cambridge University Press.

Dai, Weijia, Ginger Jin, Jungmin Lee and Michael Luca. 2018. "Aggregation of consumer ratings: an application to Yelp.com." *Quantitative Marketing and Economics* 16:289–339.

Dellarocas, Chrysanthos. 2003. "The digitization of word of mouth: Promise and challenges of online feedback mechanisms." *Management science* 49(10):1407–1424.

Dubois, Pierre and Celine Nauges. 2010. "Identifying the effect of unobserved quality and expert reviews in the pricing of experience goods: Empirical application on Bordeaux wine." *International Journal of Industrial Organization* 28(3):205 – 212.
URL: *http://www.sciencedirect.com/science/article/pii/S0167718709000848*

Ekstrand, Michael D., John T. Riedl and Joseph A. Konstan. 2011. "Collaborative filtering recommender systems." *Foundations and Trends® in Human–Computer Interaction* 4(2):81–173.

Fradkin, Andrey, Elena Grewal and David Holtz. 2018. "The determinants of online review informativeness: Evidence from field experiments on Airbnb." *Available at SSRN 2939064* .

Friberg, Richard and Erik Gronqvist. 2012. "Do Expert Reviews Affect the Demand for Wine?" *American Economic Journal: Applied Economics* 4(1):193–211.
URL: *http://www.aeaweb.org/articles?id=10.1257/app.4.1.193*

Galton, Francis. 1907. "Vox populi (the wisdom of crowds)." *Nature* 75(7):450–451.

Gergaud, Olivier, Victor Ginsburgh and Juan D Moreno-Ternero. 2021. "Wine Ratings: Seeking a Consensus among Tasters via Normalization, Approval, and Aggregation." *Journal of Wine Economics* 16(3):321–342.

Gergaud, Olivier, Victor Ginsburgh and Juan D Moreno-Ternero. 2022. "Revisiting the Judgment of Paris. The Rise and the Fall of Stag's Leap Wine Cellars.".

Ginsburgh, Victor and IsraÃ«l Zang. 2012. "Shapley Ranking of Wines." *Journal of Wine Economics* 7:169–180.

Godes, David and José C Silva. 2012. "Sequential and temporal dynamics of online opinion." *Marketing Science* 31(3):448–473.

Greene, William H. 2010. *Econometric Analysis*. Pearson Education.

Hilger, James, Greg Rafert and Sofia Villas-Boas. 2011. "Expert Opinion and the Demand for Experience Goods: An Experimental Approach in the Retail Wine Market." *The review of economics and statistics* 93:1289–1296.

Hodgson, R. and J. Cao. 2014. "Criteria for Accrediting Expert Wine Judges." *Journal of Wine Economics* 9:62–74.

Hulkower, Neal D. 2009. "The judgment of Paris according to Borda." *Journal of Wine Research* 20(3):171–182.

Lindley, Dennis V. 2006. "Analysis of a wine tasting." *Journal of Wine Economics* 1(1):33–41.

Luca, Michael. 2016. "Reviews, Reputation, and Revenue: The Case of Yelp.com." *Harvard business school* Working Paper 12-016.

Luca, Michael and Jonathan Smith. 2013. "Salience in quality disclosure: evidence from the US News college rankings." *Journal of Economics & Management Strategy* 22:58–77.

Mogstad, Magne, Joseph P Romano, Azeem Shaikh and Daniel Wilhelm. 2020. Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries. Technical report National Bureau of Economic Research.

Muchnik, Lev, Sinan Aral and Sean J Taylor. 2013. "Social influence bias: A randomized experiment." *Science* 341(6146):647–651.

Nagle, Frank and Christoph Riedl. 2014. Online word of mouth and product quality disagreement. In *Academy of management proceedings*. Number 1 Academy of Management Briarcliff Manor, NY 10510 p. 15681.

Nei, Stephen. 2017. "Frictions to Information Aggregation in Social Learning Environments." *Dissertation, Stanford University* .

Ni, Jianmo, Jiacheng Li and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fined-grained aspects. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Resnick, Paul and Rahul Sami. 2007. The influence limiter: provably manipulation-resistant recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*. ACM pp. 25–32.

Resnick, Paul and Richard Zeckhauser. 2002. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In *The Economics of the Internet and E-commerce*. Emerald Group Publishing Limited pp. 127–157.

Ricci, Francesco, Lior Rokach, Bracha Shapira and Paul B. Kantor. 2011. *Recommender Systems Handbook*. Springer New York Dordrecht Heidelberg London.

Tadelis, Steven. 2016. "Reputation and feedback systems in online platform markets." *Annual Review of Economics* 8:321–340.

Tobin, James. 1958. "Estimation of relationships for limited dependent variables." *Econometrica: journal of the Econometric Society* pp. 24–36.

White, Halbert. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: journal of the Econometric Society* pp. 817–838.

Working, Elmer J. 1927. "What do statistical "demand curves" show?" *The Quarterly Journal of Economics* 41(2):212–235.

Appendices

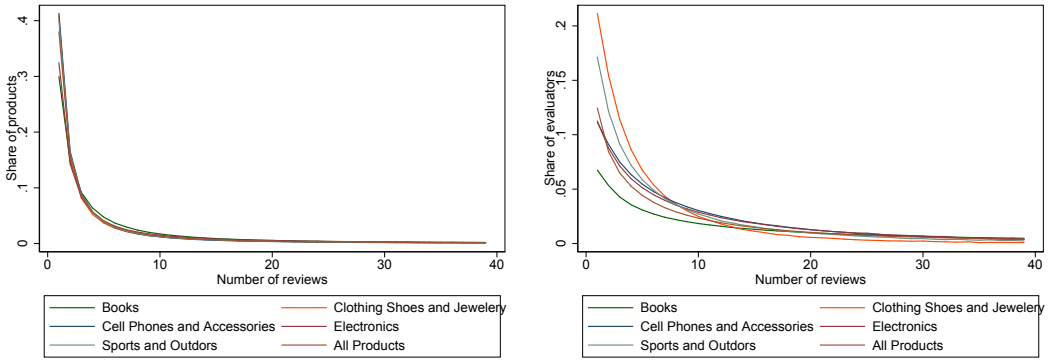
A Amazon ratings data

Table A.1: Amazon ratings per product category.

Product type	Total number of			Products					Reviewers	
				Share of products with more than			Median	Mean	with more than one rating	
	Ratings	Reviewers	Items	100 ratings	50 ratings	20 ratings	#ratings	#ratings	number	mean #ratings
All Beauty	371345	324038	32586	.02	.04	.08	2	11.40	36254	2.30
Appliances	602777	515650	30252	.04	.07	.14	2	19.93	63732	2.37
Arts Crafts and Sewing	2875917	1579230	302809	.02	.03	.07	2	9.50	477916	3.71
Automotive	7990166	3873247	925387	.01	.03	.07	2	8.63	1343949	4.06
Books	51311622	15362619	2930451	.03	.06	.14	3	17.51	6599569	6.45
CDs and Vinyl	4543369	1944316	434060	.02	.04	.09	3	1.47	627698	5.14
Cell Phones and Accessories	10063255	6211701	589534	.03	.05	.11	2	17.07	1819784	3.12
Clothing Shoes and Jewelry	32292098	12483678	2681297	.02	.04	.09	2	12.04	5541099	4.57
Digital Music	1584082	840372	456992	.00	.01	.02	1	3.47	238348	4.12
Electronics	20994353	9838676	756489	.05	.08	.17	3	27.75	3623165	4.08
Fashion	883636	749233	186189	.00	.01	.03	1	4.75	93913	2.43
Gift Cards	147194	128877	1548	.17	.27	.45	14	95.09	11555	2.59
Grocery and Gourmet Food	5074160	2695974	283507	.03	.06	.14	3	17.90	862798	3.76
Industrial and Scientific	1758333	1246131	165764	.02	.03	.08	2	1.61	262644	2.95
Luxury Beauty	574628	416174	12120	.10	.19	.37	10	47.41	91331	2.73
Movies and TV	8765568	3826085	182032	.07	.13	.23	4	48.15	1396760	4.54
Office Products	5581313	3404914	306800	.03	.06	.12	2	18.19	969642	3.24
Patio Lawn and Garden	5236058	3097405	276563	.04	.07	.14	3	18.93	928625	3.30
Prime Pantry	471614	247659	10814	.09	.20	.43	15	43.61	76104	3.94
Sports and Outdoors	12980837	6703391	957764	.02	.05	.10	2	13.55	2299429	3.73
Toys and Games	8201231	4204994	624792	.02	.05	.11	2	13.13	1406993	3.84

Notes: All Amazon data are from Ni, Li and McAuley (2019), our own computations.

Figure A.1: The distribution of the number of ratings per product (left), and of the number of ratings per user (right) for the five largest product categories (more than ten million ratings each), and for all product categories together.



B Proof of Lemma 1

To see the unbiased claims, note that for any set of true values, and realized biases and errors on all ratings, there is another set of biases and errors that have the opposite signs. That is, for each b_j and set of ε_{ij}s, consider a corresponding $\widetilde{b}_j = -b_j$, and corresponding set of $\widetilde{\varepsilon}_{ij}$s for which $\widetilde{\varepsilon}_{ij} = -\varepsilon_{ij}$. Thus, every corresponding rating $\widetilde{g}_{ij} - q_i = -(g_{ij} - q_i)$. It then follows that the corresponding estimates satisfy $\widetilde{q}_i^{one} - q_i = -(q_i^{one} - q_i)$ for each i and

that $\widetilde{b}_j^{one} = -b_j^{one}$ for each j. Then from (12) it follows that $\left(\widetilde{\sigma}_j^{one}\right)^2 = \left(\sigma_j^{one}\right)^2$ (terms are squared). It then follows that $\widetilde{q}_i^{two} - q_i = -(q_i^{two} - q_i)$ for each i and that $\widetilde{b}_j^{two} = -b_j^{two}$ for each j. Given the symmetric distributions of the b_j and set of ε_{ij}s, this implies that the distributions of the q_i^{two}s are symmetric around q_i, and the b_j^{two}s are symmetric around 0, proving the claim. ∎

C Bordeaux Wines Analysis: Reviewer's Ratings

Figure C.1: The distribution of the ratings per reviewer. Though all ratings have been re-normalized over a 100-points scale, left graph reviewers have raw ratings on a 100-points scale initially while right graph reviewers have a raw rating scale on a 20-points scale.

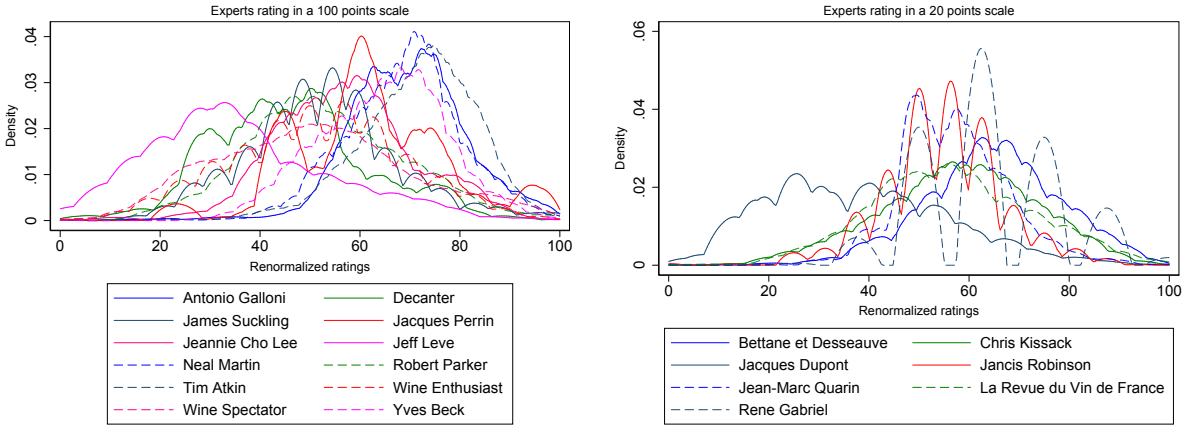


Figure C.2: Examples of a two reviewers' rescaled ratings.

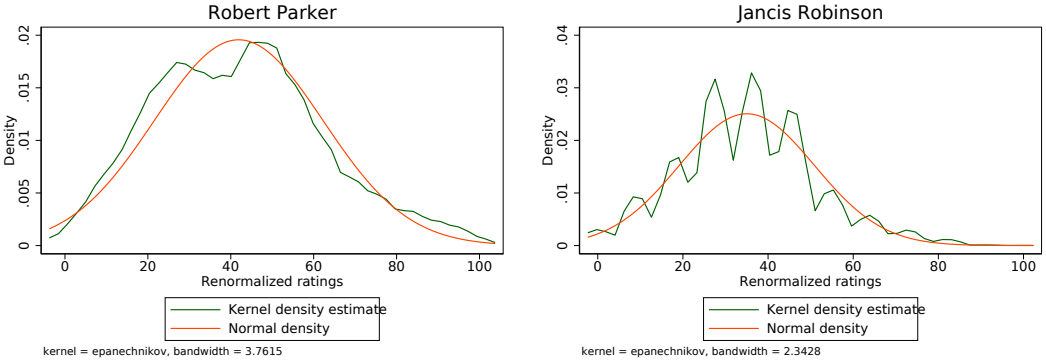


Table C.1: The top-100 rated Bordeaux wines.

Rank	q_i^{two}	Rescaled	Wine	Vintage	Type	Appellation	Classement
1	93.83	99.5	Yquem	2009	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
2	92.90	99.5	Yquem	2015	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
3	92.53	99.5	Margaux	2010	Red	Margaux	Premier Cru Classe en 1855
4	91.78	99.5	Margaux	2015	Red	Margaux	Premier Cru Classe en 1855
5	91.59	99.5	Yquem	2005	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
6	91.36	99.5	Grand Vin de Latour	2009	Red	Pauillac	Premier Cru Classe en 1855
7	91.21	99.5	Margaux	2009	Red	Margaux	Premier Cru Classe en 1855
8	91.07	99.5	Petrus	2015	Red	Pomerol	Grands Pomerol
9	90.74	99.5	Margaux	2005	Red	Margaux	Premier Cru Classe en 1855
10	90.48	99.5	Yquem	2001	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
11	90.33	99.5	Grand Vin de Latour	2010	Red	Pauillac	Premier Cru Classe en 1855
12	90.23	99.5	Grand Vin de Latour	2003	Red	Pauillac	Premier Cru Classe en 1855
13	90.02	99.5	Ausone	2015	Red	Saint Emilion Grand Cru	Premier Cru Classe A
14	89.20	99.5	Lafite Rothschild	2010	Red	Pauillac	Premier Cru Classe en 1855
15	89.14	99.5	Lafite Rothschild	2009	Red	Pauillac	Premier Cru Classe en 1855
16	88.65	99.5	Haut Brion	2009	Red	Pessac Leognan	Premier Cru Classe en 1855
17	88.54	99.5	Ausone	2005	Red	Saint Emilion Grand Cru	Premier Cru Classe A
18	88.43	99.5	La Mission Haut Brion	2000	Red	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
19	88.42	99.5	Haut Brion	2015	Red	Pessac Leognan	Premier Cru Classe en 1855
20	88.39	99	Cheval Blanc	2015	Red	Saint Emilion Grand Cru	Premier Cru Classe A
21	88.04	99	Petrus	2009	Red	Pomerol	Grands Pomerol
22	87.73	99	Lafleur	2015	Red	Pomerol	Grands Pomerol
23	87.52	99	Cheval Blanc	2010	Red	Saint Emilion Grand Cru	Premier Cru Classe A
24	87.48	99	Petrus	2010	Red	Pomerol	Grands Pomerol
25	87.47	99	Ausone	2009	Red	Saint Emilion Grand Cru	Premier Cru Classe A
26	87.26	99	Lafite Rothschild	2003	Red	Pauillac	Premier Cru Classe en 1855
27	87.21	99	Grand Vin de Latour	2005	Red	Pauillac	Premier Cru Classe en 1855
28	87.15	99	Grand Vin de Latour	2000	Red	Pauillac	Premier Cru Classe en 1855
29	86.68	99	Lafite Rothschild	2005	Red	Pauillac	Premier Cru Classe en 1855
30	86.47	99	Haut Brion	2010	Red	Pessac Leognan	Premier Cru Classe en 1855
31	86.42	99	Cheval Blanc	2009	Red	Saint Emilion Grand Cru	Premier Cru Classe A
32	85.73	99	Haut Brion	2005	Red	Pessac Leognan	Premier Cru Classe en 1855
33	85.72	99	Yquem	2014	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
34	85.52	99	Rieussec	2001	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
35	85.48	99	Leoville Las Cases	2009	Red	Saint Julien	Deuxieme Cru Classe en 1855
36	85.47	99	Lafleur	2009	Red	Pomerol	Grands Pomerol
37	85.32	99	Vieux Chateau Certan	2010	Red	Pomerol	Grands Pomerol
38	85.21	99	Mouton Rothschild	2009	Red	Pauillac	Premier Cru Classe en 1855
39	85.06	99	Grand Vin de Latour	2015	Red	Pauillac	Premier Cru Classe en 1855
40	85.05	99	Mouton Rothschild	2010	Red	Pauillac	Premier Cru Classe en 1855
41	85.05	99	Petrus	2005	Red	Pomerol	Grands Pomerol
42	85.02	99	Eglise Clinet	2009	Red	Pomerol	Grands Pomerol
43	84.98	99	Montrose	2003	Red	Saint Estephe	Deuxieme Cru Classe en 1855
44	84.61	99	Cheval Blanc	2005	Red	Saint Emilion Grand Cru	Premier Cru Classe A
45	84.59	99	Ausone	2010	Red	Saint Emilion Grand Cru	Premier Cru Classe A
46	84.43	99	Cos d'Estournel	2003	Red	Saint Estephe	Deuxieme Cru Classe en 1855
47	84.43	99	Canon	2015	Red	Saint Emilion Grand Cru	Premier Cru Classe B
48	84.30	99	Ausone	2003	Red	Saint Emilion Grand Cru	Premier Cru Classe A
49	84.18	99	La Mission Haut Brion	2015	Red	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
50	84.12	99	Mouton Rothschild	2015	Red	Pauillac	Premier Cru Classe en 1855
51	84.06	99	Suduiraut	2001	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
52	84.06	99	Lafaurie Peyraguey	2001	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
53	84.06	99	Yquem	2003	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
54	84.02	99	Yquem	2007	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
55	83.98	99	Doisy Daene. l'Extravagant	2009	Sweet	Sauternes	
56	83.96	99	Vieux Chateau Certan	2015	Red	Pomerol	Grands Pomerol
57	83.95	99	Pavie	2000	Red	Saint Emilion Grand Cru	Premier Cru Classe A
58	83.85	99	Palmer	2015	Red	Margaux	Troisieme Cru Classe en 1855
59	83.78	99	Cheval Blanc	2000	Red	Saint Emilion Grand Cru	Premier Cru Classe A
60	83.66	99	Climens	2009	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
61	83.51	99	Petrus	1998	Red	Pomerol	Grands Pomerol
62	83.42	99	Yquem	2011	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
63	83.31	99	Leoville Las Cases	2000	Red	Saint Julien	Deuxieme Cru Classe en 1855
64	83.27	99	Palmer	2009	Red	Margaux	Troisieme Cru Classe en 1855
65	83.13	98	Margaux	2003	Red	Margaux	Premier Cru Classe en 1855
66	83.09	98	Lafleur	2010	Red	Pomerol	Grands Pomerol
67	83.00	98	La Mission Haut Brion	2010	Red	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
68	82.82	98	Angelus	2015	Red	Saint Emilion Grand Cru	Premier Cru Classe A
69	82.78	98	Lafleur	2005	Red	Pomerol	Grands Pomerol
70	82.75	98	Grand Vin de Latour	2004	Red	Pauillac	Premier Cru Classe en 1855
71	82.69	98	Doisy Daene. l'Extravagant	2010	Sweet	Sauternes	
72	82.67	98	Pontet Canet	2009	Red	Pauillac	Cinquieme Cru Classe en 1855
73	82.65	98	Eglise Clinet	2010	Red	Pomerol	Grands Pomerol
74	82.59	98	Leoville Barton	2000	Red	Saint Julien	Deuxieme Cru Classe en 1855
75	82.54	98	Trotanoy	2009	Red	Pomerol	Grands Pomerol
76	82.53	98	Doisy Daene. l'Extravagant	2011	Sweet	Sauternes	
77	82.42	98	Leoville Las Cases	2005	Red	Saint Julien	Deuxieme Cru Classe en 1855
78	82.37	98	Yquem	2006	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
79	82.36	98	Doisy Daene. l'Extravagant	2005	Sweet	Sauternes	
80	82.22	98	Haut Bailly	2015	Red	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
81	82.17	98	Doisy Daene. l'Extravagant	2015	Sweet	Sauternes	
82	82.02	98	Yquem	2004	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
83	81.96	98	Figeac	2015	Red	Saint Emilion Grand Cru	Premier Cru Classe B
84	81.92	98	Petrus	2012	Red	Pomerol	Grands Pomerol
85	81.83	98	Ausone	2008	Red	Saint Emilion Grand Cru	Premier Cru Classe A
86	81.80	98	Cos d'Estournel	2010	Red	Saint Estephe	Deuxieme Cru Classe en 1855
87	81.78	97.5	Eglise Clinet	2015	Red	Pomerol	Grands Pomerol
88	81.73	97.5	Lafite Rothschild	2015	Red	Pauillac	Premier Cru Classe en 1855
89	81.64	97.5	Trotanoy	1998	Red	Pomerol	Grands Pomerol
90	81.61	97.5	Trotanoy	2015	Red	Pomerol	Grands Pomerol
91	81.60	97.5	Leoville Las Cases	2010	Red	Saint Julien	Deuxieme Cru Classe en 1855
92	81.54	97.5	Montrose	2009	Red	Saint Estephe	Deuxieme Cru Classe en 1855
93	81.52	97.5	Leoville Las Cases	2015	Red	Saint Julien	Deuxieme Cru Classe en 1855
94	81.50	97.5	Yquem	2010	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes
95	81.42	97.5	Vieux Chateau Certan	2009	Red	Pomerol	Grands Pomerol
96	81.41	97.5	Grand Vin de Latour	2014	Red	Pauillac	Premier Cru Classe en 1855
97	81.35	97.5	La Mission Haut Brion	2009	Red	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
98	81.31	97.5	Pavie	2015	Red	Saint Emilion Grand Cru	Premier Cru Classe A
99	81.20	97.5	Palmer	2010	Red	Margaux	Troisieme Cru Classe en 1855
100	81.06	97.5	Yquem	2013	Sweet	Sauternes	Premier Cru Classe en 1855 - Sauternes

Online Supplementary Appendices (D–G) of the paper:

Finding the Wise and the Wisdom in a Crowd: Estimating Underlying Qualities of Reviewers and Items

by Nicolas Carayol and Matthew O. Jackson

D Additional Analysis of Bordeaux Wines and Experts' Ratings

D.1 More on the Data and Estimations

Table D.1: The Bordeaux Wines, by Appellation.

Appellation	Number of wines/vintages	Number of ratings
Barsac	17	154
Blaye	4	17
Bordeaux	140	797
Bordeaux Superieur	42	185
Canon Fronsac	12	60
Castillon Cotes de Bordeaux	3	21
Cotes de Blaye	3	9
Cotes de Bourg	15	64
Cotes de Castillon	65	407
Cotes de Franc	13	99
Entre deux mers	10	35
Fronsac	71	350
Graves	134	536
Haut Medoc	308	1,839
Lalande de Pomerol	86	494
Listrac Medoc	70	429
Lussac Saint Emilion	14	38
Margaux	512	4,128
Medoc	135	639
Montagne Saint Emilion	16	57
Moulis en Medoc	65	449
Pauillac	447	3,923
Pessac Leognan	448	3,635
Pessac Leognan, Blanc	288	2,369
Pomerol	657	4,763
Premieres Cotes de Blaye	5	19
Premieres Cotes de Bordeaux	39	157
Puisseguin Saint Emilion	12	62
Saint Emilion	470	2,426
Saint Emilion Grand Cru	1,183	8,733
Saint Estephe	283	2,239
Saint Georges Saint Emilion	1	2
Saint Julien	307	2,677
Sainte Foy Bordeaux	5	35
Sauternes	465	3,594
Vin de France	1	5

Table D.2: Bordeaux Wines and ratings, by Official Rankings.

Classement (official ranking)	Number of wines/vintages	Number of ratings
Cinquieme Cru Classe en 1855	316	2,864
Deuxieme Cru Classe en 1855	242	2,338
Deuxieme Cru Classe en 1855 - Sauternes	191	1,509
Grand Cru Assimile-Medoc	313	2,424
Grand Cru Classe de Graves (Blanc)	113	985
Grand Cru Classe de Graves (Rouge)	204	1,862
Grand Cru Classe de St Emilion	862	5,839
Grands Pomerol	346	3,002
Premier Cru Classe A	72	673
Premier Cru Classe B	236	2,176
Premier Cru Classe en 1855	90	871
Premier Cru Classe en 1855 - Sauternes	187	1,654
Quatrieme Cru Classe en 1855	168	1,537
Seconds Vins	195	1,624
Troisieme Cru Classe en 1855	231	2,094

Figure D.1: Selection: how reviewers select the wines they evaluate (with respect to estimated quality).

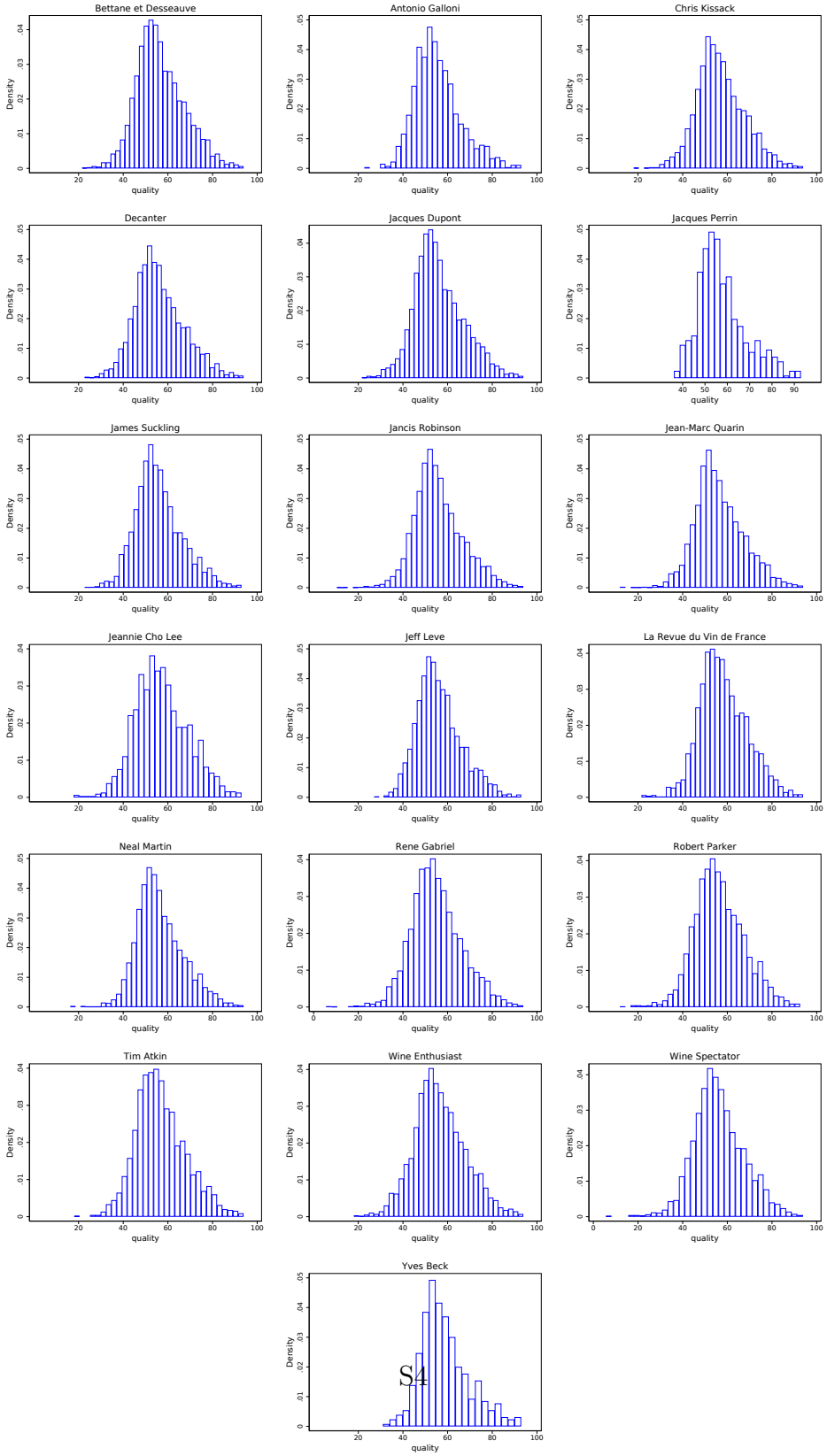


Figure D.2: Time distribution of ratings and wine/vintages.

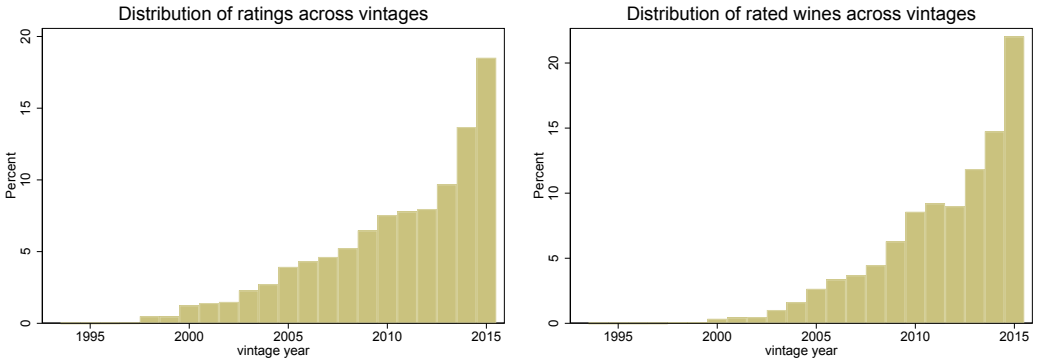


Figure D.3: The biases of the experts.

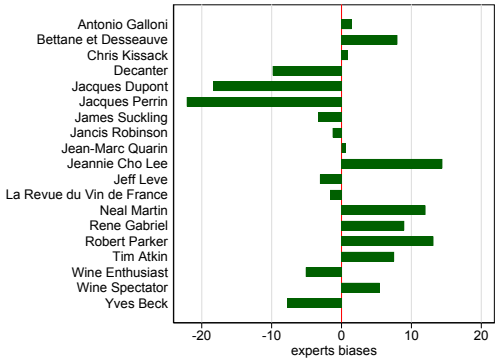


Figure D.4: The accuracies of the experts (left graph) and the correlation of their ratings with the estimated wine qualities (right graph).

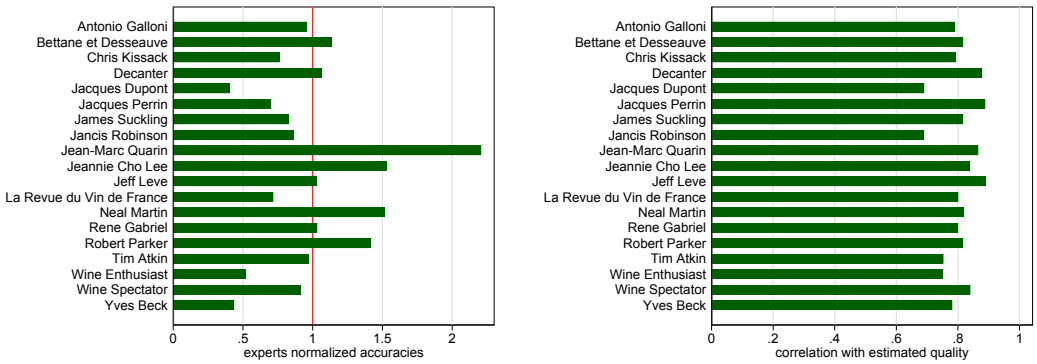


Figure D.5: The relationship between the accuracies of experts and the correlation of their ratings with the estimated wine qualities.

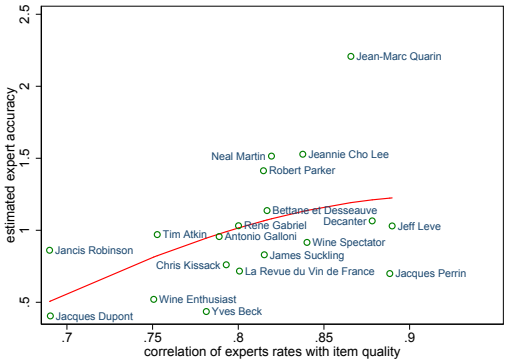
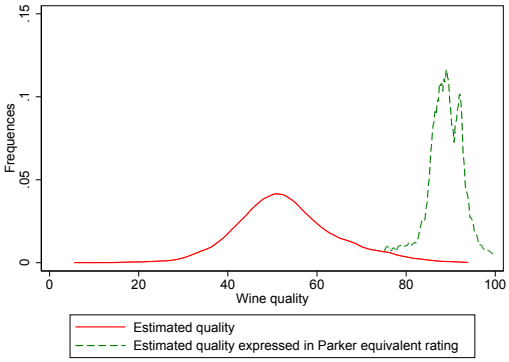


Figure D.6: The rescaling of our estimated qualities to adopt the "Parker scale".



D.2 Monte Carlo Simulations Calibrated on Bordeaux Wine Data

We consider another measurement here that we call "*fitness*". It is the share of the per-item average error in the data that is resolved by our estimation:

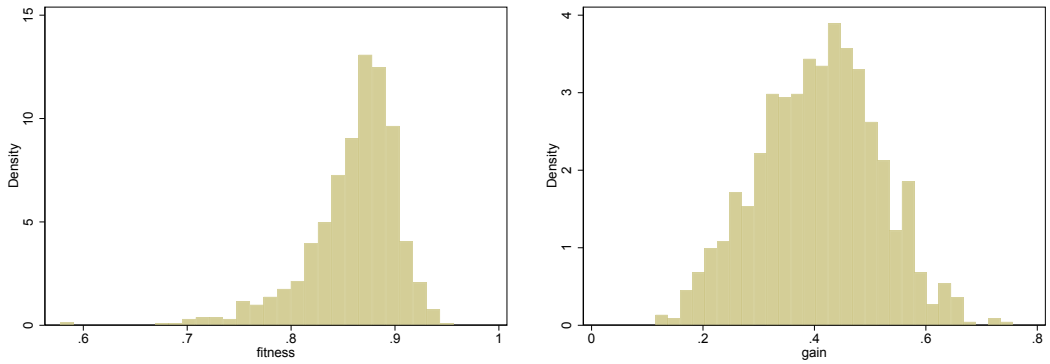
$$\text{Fitness} = 1 - \frac{E\left[(q_i^{two} - q_i)^2\right]}{E\left[(g_{ij} - q_i)^2\right]} = 1 - \frac{E\left[(q_i^{two} - q_i)^2\right]}{E\left[(b_i + \varepsilon_{ij})^2\right]}. \tag{22}$$

Table D.3: Descriptive statistics of Fitness and Gain calculated on 1,000 Monte Carlo numerical experiments calibrated on Bordeaux Wine Data

Stats	Fitness	Gain
mean	.82	.26
median	.84	.26
sd	.05	.12

Notes: Bordeaux wine rating calibration: $r = 38,279$, $m = 19$, $n = 5,371$, $\underline{q} = 8.862$, $\overline{q} = 95.501$, $\sigma_b = 9.765$, $\underline{\sigma} = 5.630$, and $\overline{\sigma} = 12.914$.

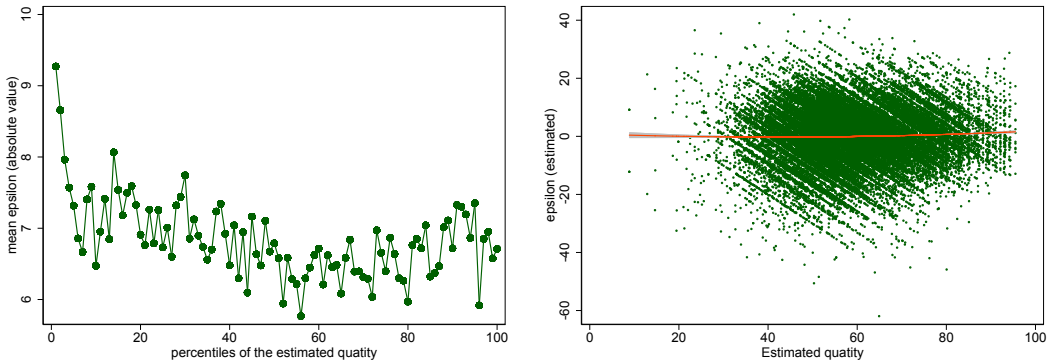
Figure D.7: Histograms of Fitness and Gain statistics calculated on 1,000 Monte Carlo numerical experiments calibrated on Bordeaux Wine Data



Notes: Bordeaux wine rating calibration: $r = 38,279$, $m = 19$, $n = 5,371$, $\underline{q} = 8.862$, $\overline{q} = 95.501$, $\sigma_b = 9.765$, $\underline{\sigma} = 5.630$, and $\overline{\sigma} = 12.914$.

D.3 Errors and Quality for Bordeaux Wines

Figure D.8: The relation between percentiles of (estimated) quality (corrected from the expert bias) and experts' (estimated) errors.



D.4 Experts Accuracies on and Quality Ranking of Bordeaux Red Wines

Table D.4: Experts accuracies and biases ranking of Bordeaux red wines only.

Expert	$\left(\sigma_j^{two}\right)^2$	A^{two}	Corr (g_{ij}, q_i^{two})	b_j^{two}	n_j
Antonio Galloni	73.61	0.91	0.78	1.54	991
Bettane et Desseauve	61.32	1.18	0.82	8.28	2,574
Chris Kissack	90.06	0.72	0.79	0.21	1,983
Decanter	62.02	1.09	0.88	-10.24	1,961
Jacques Dupont	142.46	0.41	0.72	-17.35	2,600
Jacques Perrin	89.21	0.70	0.88	-22.56	427
James Suckling	77.76	0.85	0.82	-3.49	1,722
Jancis Robinson	80.32	0.84	0.68	-1.29	3,100
Jean-Marc Quarin	35.34	2.52	0.89	1.28	2,473
Jeannie Cho Lee	52.99	1.39	0.83	14.67	1,050
Jeff Leve	67.80	0.99	0.88	-3.32	1,408
La Revue du Vin de France	83.74	0.77	0.81	-1.30	1,814
Neal Martin	52.52	1.45	0.82	12.02	2,457
Rene Gabriel	67.03	1.07	0.80	8.78	4,058
Robert Parker	57.25	1.34	0.81	13.06	2,547
Tim Atkin	76.34	0.89	0.75	7.56	1,583
Wine Enthusiast	112.54	0.54	0.77	-5.21	2,050
Wine Spectator	74.57	0.88	0.84	5.08	3,087
Yves Beck	135.35	0.45	0.78	-7.72	394

Figure D.9: The accuracies of the experts (left graph) and the correlation of their ratings with the estimated red wine qualities (right graph).

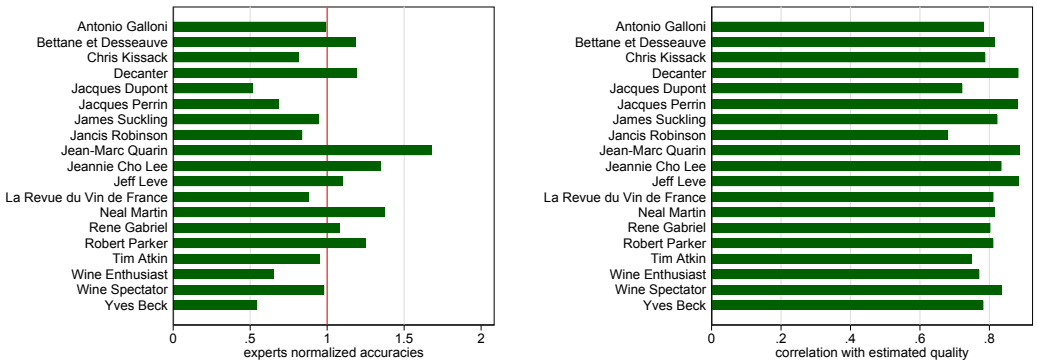


Table D.5: The top-100 rated Bordeaux red wines.

Rank	q_i^{two}	Rescaled	Wine	Vintage	Appellation	Classement
1	92,69	99,5	Margaux	2010	Margaux	Premier Cru Classe en 1855
2	91,82	99,5	Margaux	2015	Margaux	Premier Cru Classe en 1855
3	91,31	99,5	Grand Vin de Latour	2009	Pauillac	Premier Cru Classe en 1855
4	91,28	99,5	Margaux	2009	Margaux	Premier Cru Classe en 1855
5	91,04	99,5	Petrus	2015	Pomerol	Grands Pomerol
6	90,74	99,5	Margaux	2005	Margaux	Premier Cru Classe en 1855
7	90,50	99,5	Grand Vin de Latour	2010	Pauillac	Premier Cru Classe en 1855
8	90,37	99,5	Grand Vin de Latour	2003	Pauillac	Premier Cru Classe en 1855
9	90,04	99,5	Ausone	2015	Saint Emilion Grand Cru	Premier Cru Classe A
10	89,25	99,5	Lafite Rothschild	2010	Pauillac	Premier Cru Classe en 1855
11	89,00	99,5	Lafite Rothschild	2009	Pauillac	Premier Cru Classe en 1855
12	88,75	99,5	Haut Brion	2009	Pessac Leognan	Premier Cru Classe en 1855
13	88,68	99,5	La Mission Haut Brion	2000	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
14	88,46	99,5	Ausone	2005	Saint Emilion Grand Cru	Premier Cru Classe A
15	88,45	99,5	Haut Brion	2015	Pessac Leognan	Premier Cru Classe en 1855
16	88,42	99,0	Cheval Blanc	2015	Saint Emilion Grand Cru	Premier Cru Classe A
17	88,12	99,0	Petrus	2009	Pomerol	Grands Pomerol
18	87,70	99,0	Lafleur	2015	Pomerol	Grands Pomerol
19	87,66	99,0	Cheval Blanc	2010	Saint Emilion Grand Cru	Premier Cru Classe A
20	87,59	99,0	Ausone	2009	Saint Emilion Grand Cru	Premier Cru Classe A
21	87,51	99,0	Petrus	2010	Pomerol	Grands Pomerol
22	87,43	99,0	Grand Vin de Latour	2000	Pauillac	Premier Cru Classe en 1855
23	87,35	99,0	Lafite Rothschild	2003	Pauillac	Premier Cru Classe en 1855
24	86,92	99,0	Grand Vin de Latour	2005	Pauillac	Premier Cru Classe en 1855
25	86,67	99,0	Lafite Rothschild	2005	Pauillac	Premier Cru Classe en 1855
26	86,42	99,0	Cheval Blanc	2009	Saint Emilion Grand Cru	Premier Cru Classe A
27	86,41	99,0	Haut Brion	2010	Pessac Leognan	Premier Cru Classe en 1855
28	85,43	99,0	Leoville Las Cases	2009	Saint Julien	Deuxieme Cru Classe en 1855
29	85,41	99,0	Haut Brion	2005	Pessac Leognan	Premier Cru Classe en 1855
30	85,37	99,0	Lafleur	2009	Pomerol	Grands Pomerol
31	85,30	99,0	Vieux Chateau Certan	2010	Pomerol	Grands Pomerol
32	85,26	99,0	Mouton Rothschild	2009	Pauillac	Premier Cru Classe en 1855
33	85,13	99,0	Grand Vin de Latour	2015	Pauillac	Premier Cru Classe en 1855
34	85,09	99,0	Mouton Rothschild	2010	Pauillac	Premier Cru Classe en 1855
35	85,06	99,0	Eglise Clinet	2009	Pomerol	Grands Pomerol
36	84,98	99,0	Montrose	2003	Saint Estephe	Deuxieme Cru Classe en 1855
37	84,94	99,0	Petrus	2005	Pomerol	Grands Pomerol
38	84,56	99,0	Cos d'Estournel	2003	Saint Estephe	Deuxieme Cru Classe en 1855
39	84,55	99,0	Ausone	2010	Saint Emilion Grand Cru	Premier Cru Classe A
40	84,44	99,0	Canon	2015	Saint Emilion Grand Cru	Premier Cru Classe B
41	84,44	99,0	Cheval Blanc	2005	Saint Emilion Grand Cru	Premier Cru Classe A
42	84,31	99,0	Pavie	2000	Saint Emilion Grand Cru	Premier Cru Classe A
43	84,28	99,0	Ausone	2003	Saint Emilion Grand Cru	Premier Cru Classe A
44	84,24	99,0	La Mission Haut Brion	2015	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
45	84,15	99,0	Mouton Rothschild	2015	Pauillac	Premier Cru Classe en 1855
46	83,89	99,0	Palmer	2015	Margaux	Troisieme Cru Classe en 1855
47	83,81	99,0	Vieux Chateau Certan	2015	Pomerol	Grands Pomerol
48	83,75	99,0	Petrus	1998	Pomerol	Grands Pomerol
49	83,69	99,0	Cheval Blanc	2000	Saint Emilion Grand Cru	Premier Cru Classe A
50	83,68	99,0	Leoville Las Cases	2000	Saint Julien	Deuxieme Cru Classe en 1855
51	83,37	98,0	Palmer	2009	Margaux	Troisieme Cru Classe en 1855
52	83,19	98,0	Margaux	2003	Margaux	Premier Cru Classe en 1855
53	83,12	98,0	Lafleur	2010	Pomerol	Grands Pomerol
54	83,02	98,0	La Mission Haut Brion	2010	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
55	82,90	98,0	Angelus	2015	Saint Emilion Grand Cru	Premier Cru Classe A
56	82,74	98,0	Leoville Barton	2000	Saint Julien	Deuxieme Cru Classe en 1855
57	82,66	98,0	Eglise Clinet	2010	Pomerol	Grands Pomerol
58	82,59	98,0	Lafleur	2005	Pomerol	Grands Pomerol
59	82,58	98,0	Pontet Canet	2009	Pauillac	Cinquieme Cru Classe en 1855
60	82,52	98,0	Grand Vin de Latour	2004	Pauillac	Premier Cru Classe en 1855
61	82,41	98,0	Leoville Las Cases	2005	Saint Julien	Deuxieme Cru Classe en 1855
62	82,31	98,0	Trotanoy	2009	Pomerol	Grands Pomerol
63	82,16	98,0	Haut Bailly	2015	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
64	82,00	98,0	Cos d'Estournel	2010	Saint Estephe	Deuxieme Cru Classe en 1855
65	81,95	98,0	Figeac	2015	Saint Emilion Grand Cru	Premier Cru Classe B
66	81,93	98,0	Leoville Las Cases	2010	Saint Julien	Deuxieme Cru Classe en 1855
67	81,89	98,0	Trotanoy	1998	Pomerol	Grands Pomerol
68	81,87	97,5	Petrus	2012	Pomerol	Grands Pomerol
69	81,80	97,5	Eglise Clinet	2015	Pomerol	Grands Pomerol
70	81,72	97,5	Lafite Rothschild	2015	Pauillac	Premier Cru Classe en 1855
71	81,66	97,5	Ausone	2008	Saint Emilion Grand Cru	Premier Cru Classe A
72	81,61	97,5	Trotanoy	2015	Pomerol	Grands Pomerol
73	81,54	97,5	Leoville Las Cases	2015	Saint Julien	Deuxieme Cru Classe en 1855
74	81,44	97,5	Pavie	2015	Saint Emilion Grand Cru	Premier Cru Classe A
75	81,42	97,5	Montrose	2009	Saint Estephe	Deuxieme Cru Classe en 1855
76	81,39	97,5	Grand Vin de Latour	2014	Pauillac	Premier Cru Classe en 1855
77	81,39	97,5	Vieux Chateau Certan	2009	Pomerol	Grands Pomerol
78	81,34	97,5	La Mission Haut Brion	2009	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
79	81,26	97,5	Palmer	2010	Margaux	Troisieme Cru Classe en 1855
80	81,17	97,5	Ducru Beaucaillou	2015	Saint Julien	Deuxieme Cru Classe en 1855
81	80,96	97,5	Troplong Mondot	2005	Saint Emilion Grand Cru	Premier Cru Classe B
82	80,92	97,5	Ducru Beaucaillou	2009	Saint Julien	Deuxieme Cru Classe en 1855
83	80,55	97,5	Ducru Beaucaillou	2010	Saint Julien	Deuxieme Cru Classe en 1855
84	80,52	97,5	Palmer	2005	Margaux	Troisieme Cru Classe en 1855
85	80,52	97,5	Trotanoy	2010	Pomerol	Grands Pomerol
86	80,46	97,5	Cos d'Estournel	2005	Saint Estephe	Deuxieme Cru Classe en 1855
87	80,45	97,5	Tertre Roteboeuf	2015	Saint Emilion	
88	80,37	97,5	Mouton Rothschild	2006	Pauillac	Premier Cru Classe en 1855
89	80,24	97,5	Mouton Rothschild	2002	Pauillac	Premier Cru Classe en 1855
90	80,24	97,5	Lafite Rothschild	2000	Pauillac	Premier Cru Classe en 1855
91	80,04	97,5	Vieux Chateau Certan	1998	Pomerol	Grands Pomerol
92	80,04	97,5	Haut Brion	1998	Pessac Leognan	Premier Cru Classe en 1855
93	79,99	97,5	Pontet Canet	2010	Pauillac	Cinquieme Cru Classe en 1855
94	79,98	97,5	Leoville Las Cases	2006	Saint Julien	Deuxieme Cru Classe en 1855
95	79,92	97,5	Pichon Baron	2010	Pauillac	Deuxieme Cru Classe en 1855
96	79,79	97,5	Margaux	2006	Margaux	Premier Cru Classe en 1855
97	79,61	97,5	Margaux	2000	Margaux	Premier Cru Classe en 1855
98	79,61	97,5	Lynch Bages	2000	Pauillac	Cinquieme Cru Classe en 1855
99	79,61	97,5	Evangile	2000	Pomerol	Grands Pomerol
100	79,58	97,5	Le Pin	2010	Pomerol	Grands Pomerol

D.5 Rating Left-Bank Versus Right-Bank Bordeaux Red Wines

Figure D.10: The relationship between the differences in accuracies and the differences in biases (between left bank and right bank wines).

E More on Prices and Ratings of Bordeaux Wines

Table E.1: Markets surveyed, stores and prices.

Market	Number of stores	Number of wines	Number of prices
Hong Kong	222	6,502	13,368
New York	342	7,305	12,052
Paris	354	10,537	17,887

Figure E.1: Prices in the three markets (in local currency).

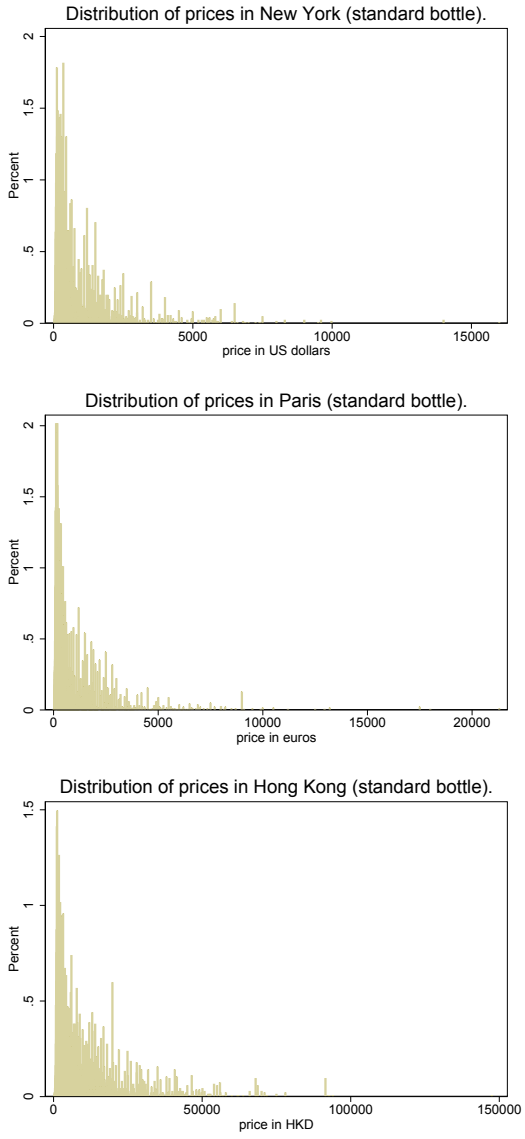


Table E.2: Top-20 most surveyed stores (restaurants).

Store	Market	Number of Wines
L'Atelier de Joel Robuchon - HK	Hong Kong	505
La Truffiere	Paris	452
Le Cinq - Paris	Paris	308
Le Pre Catelan	Paris	300
Apicius	Paris	291
Le Carre des Feuillants	Paris	289
Petrus - HK	Hong Kong	258
Epicure	Paris	251
Cepage	Hong Kong	238
L Abeille (Shangri-La)	Paris	199
KO Dining Group (Messina, Yu Lei, Kazuo Okuda)	Hong Kong	190
Per Se	New York	190
Mandarin Oriental Paris - Sur Mesure, Camelia	Paris	183
21 Club	New York	180
Shang Palace (Shangri-La) - Paris	Paris	163
Le Meurice	Paris	163
Au Trou Gascon	Paris	153
Spoon	Hong Kong	149
Alain Ducasse au Plaza Athenee	Paris	148
The Steak House winebar + grill	Hong Kong	144

Table E.3: Retail prices as a function of estimated wine quality and of salient and best en primeur ratings. Without numerous fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Estimated quality	3.448^{+}		3.357^{+}	3.531^{+}	3.538^{+}	3.396^{+}	3.885^{+}	2.161^{+}	3.007^{+}
	(24.05)		(23.74)	(14.64)	(14.45)	(21.74)	(14.86)	(5.65)	(6.03)
Best rating				-0.235			$-0.427^{\#}$	-0.346	-0.237
				(-0.97)			(-1.99)	(-1.63)	(-0.98)
R. Parker rating					0.243				
					(0.89)				
J. Robinson rating						$0.328^{\star}$			
						(2.94)			
Average rating								1.703^{+}	0.569
								(4.76)	(1.42)
N	43307	43307	43307	43307	36109	28774	22264	22264	5090
r2	0.601	0.204	0.604	0.604	0.587	0.643	0.593	0.597	0.555
aic	85915.5	115861.6	85589.8	85568.2	71797.5	51826.6	39623.4	39409.6	8494.6
bic	85932.9	115870.2	85615.9	85602.9	71831.5	51859.7	39655.5	39449.7	8527.3

Notes: t-statistics are in parentheses. The standard errors are two-way clustered at the wine×vintage level and at the store level. Significance levels: $^{\#}p < 0.1$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All ratings are corrected to span a 0-100 scale (see Equation 15). The price variable and the listed variables are all in logs so that coefficients can be interpreted as elasticities. All regressions but Column 2 include the variance of the considered wine ratings. All models include a limited number of fixed effects: year, type (red, white or sweet), and vintage fixed effects. Columns 7 and 8 consider only the prices of wines formed less than 10 years after production (thus 9 years after the "en primeur" reviews), and Column 9, only the prices of wines formed less than 5 years after production.

Table E.4: Retail prices as a function of each expert "en primeur" ratings.

Expert	coef	t-stat	N	R2	AIC	BIC
Antonio Galloni	3.318	(0.39)	79	0.947	-58.31	-48.83
Bettane et Desseauve	1.707^{+}	(12.12)	21664	0.756	29388.7	30825.8
Chris Kissack	0.884^{+}	(5.77)	10059	0.730	12120.5	13267.8
Decanter	1.155^{+}	(9.48)	4241	0.794	4362.7	5023.3
Jacques Dupont	0.541^{+}	(8.64)	23898	0.753	33457.4	35106.1
James Suckling	1.378^{+}	(6.86)	2080	0.822	1866.2	2294.8
Jancis Robinson	1.198^{+}	(13.43)	28774	0.766	40215.6	42232.8
Jean-Marc Quarin	2.536^{+}	(13.97)	21645	0.779	27484.4	29041.0
Jeannie Cho Lee	3.884^{+}	(6.20)	1119	0.882	865.8	1177.1
Jeff Leve	1.172^{+}	(5.52)	1066	0.850	606.5	924.7
La Revue du Vin de France	1.057^{+}	(9.02)	14118	0.771	17906.8	19115.7
Neal Martin	1.119^{+}	(5.54)	15523	0.772	19186.7	20548.4
Rene Gabriel	1.396^{+}	(13.83)	37285	0.756	55249.5	57585.7
Robert Parker	1.956^{+}	(7.54)	36109	0.755	53433.0	55624.5
Tim Atkin	1.071^{+}	(6.31)	3304	0.811	3627.5	4250.0
Wine Enthusiast	0.994^{+}	(11.56)	16636	0.792	22813.1	24125.4
Wine Spectator	1.170^{+}	(14.01)	38917	0.751	58696.9	61027.7

Notes: Column "coef" exhibits estimated coefficients of each expert ratings in a linear regression on wines retail prices. In all regressions, standard errors are clustered at the wine×vintage level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include rating year, vintage×appellation, official ranking, type (color), and retail shop fixed effects. Ratings are corrected to span the 0-100 scale (see Equation 15). Prices and ratings are in log so that coefficients can be interpreted as elasticities. Regressions did not converge for Jacques Perrin and Yves Beck. Some did converge, but the expert rated a limited number of wines that are priced (less than 1,200), like Antonio Galloni, Jeff Leve and Jeannie Cho Lee. Regression results are reported here but not in Figure 4.

Table E.5: Retail prices as a function of "en primeur" ratings by the top-5 most influential experts (on prices). All markets.

	(1)	(2)	(3)	(4)
Jean-Marc Quarin	2.027^{+}	1.660^{+}	1.552^{+}	1.298^{+}
	(8.56)	(7.73)	(7.49)	(6.58)
Robert Parker	$1.079^{\#}$	0.783	0.647	0.749
	(2.44)	(1.75)	(1.47)	(1.65)
Bettane et Desseauve		0.986^{+}	0.869^{+}	0.704^{+}
		(6.41)	(5.91)	(4.77)
Rene Gabriel			$0.479^{\star}$	0.281
			(3.14)	(1.90)
Jancis Robinson				0.734^{+}
				(8.39)
N	17766	16900	16780	16572
r2	0.792	0.800	0.803	0.816
aic	21898.0	20359.0	19956.3	18534.9
bic	23151.4	21519.3	21107.8	19684.5

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include vintage, rating year, vintage×appellation, type (color), and official ranking. Ratings are corrected to span the 0-100 scale (see Equation 15). Prices and ratings are in log so that coefficients can be interpreted as elasticities.

Table E.6: Retail prices as a function of "en primeur" ratings by the top-5 most influential experts (on prices). Paris market.

	(1)	(2)	(3)	(4)
Jean-Marc Quarin	2.097^{+}	1.706^{+}	1.616^{+}	1.436^{+}
	(7.58)	(6.96)	(7.03)	(6.71)
Robert Parker	0.639	0.354	0.232	0.251
	(1.32)	(0.73)	(0.49)	(0.53)
Bettane et Desseauve		1.020^{+}	0.895^{+}	0.785^{+}
		(5.68)	(5.27)	(4.83)
Rene Gabriel			$0.481^{\star}$	$0.313^{\#}$
			(3.06)	(2.07)
Jancis Robinson				0.648^{+}
				(6.74)
N	8083	7597	7529	7450
r2	0.806	0.815	0.818	0.829
aic	9433.4	8647.0	8441.9	7911.0
bic	10490.0	9638.7	9425.5	8886.2

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include vintage, rating year, vintage×appellation and official ranking. Ratings are corrected to span the 0-100 scale (see Equation 15). Prices and ratings are in log so that coefficients can be interpreted as elasticities.

Table E.7: Retail prices as a function of "en primeur" ratings by the top-5 most influential experts (on prices). New York market.

	(1)	(2)	(3)	(4)
Jean-Marc Quarin	1.562^{+}	1.293^{+}	1.206^{+}	0.890^{+}
	(8.84)	(7.30)	(6.81)	(5.45)
Robert Parker	1.994^{+}	1.759^{+}	1.572^{+}	1.758^{+}
	(9.01)	(7.94)	(6.75)	(7.87)
Bettane et Desseauve		0.715^{+}	0.617^{+}	$0.471^{\star}$
		(4.72)	(3.98)	(3.05)
Rene Gabriel			$0.406^{\star}$	0.212
			(2.65)	(1.50)
Jancis Robinson				0.698^{+}
				(7.88)
N	4579	4408	4381	4333
r2	0.844	0.849	0.851	0.861
aic	4002.5	3749.1	3651.0	3328.2
bic	4799.7	4528.8	4430.0	4112.2

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include vintage, rating year, vintage×appellation and official ranking. Ratings are corrected to span the $0-100$ scale (see Equation 15). Prices and ratings are in log so that coefficients can be interpreted as elasticities.

Table E.8: Retail prices as a function of "en primeur" ratings by the top-5 most influential experts (on prices). Hong Kong market.

	(1)	(2)	(3)	(4)
Jean-Marc Quarin	2.132^{+}	1.786^{+}	1.645^{+}	1.253^{+}
	(7.06)	(6.12)	(5.68)	(4.41)
Robert Parker	$1.451^{\#}$	1.060	0.951	1.234
	(2.48)	(1.77)	(1.61)	(1.94)
Bettane et Desseauve		1.080^{+}	0.973^{+}	$0.647^{\star}$
		(5.21)	(4.88)	(3.07)
Rene Gabriel			$0.504^{\#}$	0.273
			(2.32)	(1.30)
Jancis Robinson				0.906^{+}
				(6.49)
N	5104	4895	4870	4789
r2	0.763	0.769	0.772	0.792
aic	7223.5	6821.1	6731.6	6179.2
bic	7994.9	7561.6	7484.5	6930.2

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include vintage, rating year, vintage×appellation and official ranking. Ratings are corrected to span the 0-100 scale (see Equation 15). Prices and ratings are in log so that coefficients can be interpreted as elasticities.

F Estimated Qualities and the Re-Rating of Bordeaux Wines

F.1 Re-Rating Data and Estimation Strategy

Rerating data of the same exact wines/vintages are available for six experts: Decanter, James Suckling, Jancis Robinson, Neal Martin, Robert Parker, and Wine Spectator. That makes a total of 12,739 revised ratings that follow an initial "en primeur" rating (examined in Section 5.1) of 2,977 distinct wine/vintages by the same experts. Table F.1 in Online Appendix F.3 provides more information on the re-rating data by experts. Decanter re-rated only a few vines whereas Jancis Robinson, Robert Parker, and Wine Spectator re-rated more than two thousand wine/vintages. Jancis Robinson re-rates each of those wines in average 2.5 times whereas Robert Parker does so 1.5 times. The average intial rating of those wines is 62 and the average adjustment (the difference between the re-rating and the initial rate) is 12.8, which is pretty large.

When re-rating a wine that an expert already rated in the past, her/his new rating may be correlated to her/his own initial rating for several reasons. The expert has specific tastes and re-rating will basically be correlated with the initial rating because of that bias. The expert could also remember the initial rating and also take this first "signal" into account. She/he may also wish to minimally deviate from the initial rating (for consistency or to avoid signaling a limited accuracy). As these initial ratings are also correlated with unobserved quality, we therefore include the initial rating as a control. Moreover, other ratings, that are also correlated with the unobserved quality, may influence experts. We thus control for the salient experts ratings (Parker and Robinson), as well as the best rating. Each time the ratings of some expert are controlled for (for example Robert Parker's), the re-ratings of that expert cannot be considered as well as the wines they did not rate, and thus this comes at the cost of available data. When the best rating of each wine is used as a control, then the expert rating of that expert for this wine in not considered as well.

Different wines age in different ways. We thus also include numerous fixed effects that account for the evolution of the quality of the wine over the years: re-rating year, aging, All regressions include aging, vintage×appellation, type (color), and official ranking. As different experts may re-rate wines in different ways, we also include expert fixed effects.

Section G.2 of this Online Appendix provides some micro-foundations for re-rating that are consistent with our empirical findings. When re-rating a wine, the expert considers his or her initial rating as well as a new signal (tasting) and may be influenced by some other expert.

Table F.1: Summary statistics on the re-rating data, by expert.

Expert	# Re-ratings	Number of wines	Mean initial rating	Adjustment (share)
Decanter	6	5	39.17	0.40
James Suckling	585	499	57.81	0.24
Jancis Robinson	5,173	2,074	56.46	0.21
Neal Martin	1,380	824	71.86	0.18
Robert Parker	3,307	2,105	72.12	0.17
Wine Spectator	2,288	2,157	63.19	0.38

F.2 Results

The results are presented in Table F.2. In the first column, the rerating is regressed only on estimated quality, on the top of all fixed effects. In Columns 2-4, the other salient ratings are introduced one at the time, and altogether in Column 5. Standard errors are clustered to account for potential correlation between observations at the wine/vintage/expert level.

Our estimate of quality is, in all regressions, a very significant predictor of the decisions that experts make in their rating (always significant at the .001 level), even with the many fixed effects introduced and controlling for salient experts ratings. Experts' are consistently adjusting their ratings to be closer to our estimated quality. The coefficients are large (from .169 to .236), and close to the initial rating coefficients (from 0.161 to 0.234). As all those variables are in logs, the coefficients can again be interpreted as elasticities. According to Column 5, our preferred specification, a 10 percent increase in the estimated quality raises the new rating by 2.1 percent whereas a similar increase in the previous rating of the same expert only raises the re-rating by 1.9 percent. Robert Parker's ratings of "en primeur" wines do not correlate with the re-ratings, nor do best ratings. Jancis Robinson's ratings significantly, but slightly negatively, correlate with re-ratings.

F.3 An Alternative Empirical Strategy for Estimating Re-Ratings of Bordeaux Wines (in Differences)

In Table F.4, we also examine how experts' changes in ratings (or rating adjustments) depend on the difference between our estimated quality and their initial rating. We call that difference the "theoretical adjustment" which is also net of each experts' bias. The ratings are not in logs in these regressions so as to be able to compare the initial error and initial rating scales (similar results hold with log ratings). All other controls used in the previous regression remain. These regressions show that experts adjust their ratings about 21 to 30

Table F.2: Re-rating as a function of estimated quality, of en primeur rating by the same expert, and of the "salient" best en primeur rating.

	(1)	(2)	(3)	(4)	(5)
Estimated quality	0.157^{+}	0.174^{+}	0.179^{+}	0.226^{+}	0.203^{+}
	(14.68)	(7.20)	(8.57)	(13.28)	(4.52)
En primeur initial rating	0.235^{+}	0.226^{+}	0.213^{+}	0.160^{+}	0.193^{+}
	(25.00)	(20.21)	(17.87)	(11.41)	(8.98)
Best rating		0.0174			-0.0116
		(0.78)			(-0.31)
R. Parker rating			0.0213		0.0236
			(1.27)		(1.34)
J. Robinson rating				$-0.0197^{\star}$	$-0.0251^{\#}$
				(-3.27)	(-1.98)
N	12739	10426	6958	5260	2147
r2	0.738	0.723	0.704	0.714	0.734
aic	-23988.2	-18511.4	-12190.8	-14083.0	-5200.9
bic	-23705.0	-18228.6	-11937.5	-13840.0	-5002.4

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage×expert level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include aging, vintage×appellation, official ranking, type (color), re-rating year, and expert fixed effects. Ratings are corrected to span the 0-100 scale (see Equation 15). All ratings are in log.

percent (depending on the specification) in the direction that corrects their initial error with respect to the estimated quality of "en primeur" wines. Also, adjustments move against the initial rating so that if the initial rating was high, it is likely that the difference will be small, more likely negative.

Table F.3: Re-rating as a function of estimated quality, of en primeur rating by the same expert, and of the "salient" best en primeur rating. Only re-ratings published more than two years since the initial en primeur rating.

	(1)	(2)	(3)	(4)	(5)
Estimated quality	0.302^{+}	0.353^{+}	0.340^{+}	0.274^{+}	0.204^{+}
	(12.06)	(6.60)	(8.71)	(9.38)	(3.55)
En primeur initial rating	0.0924^{+}	0.0649^{+}	$0.0453^{\#}$	0.126^{+}	0.130^{+}
	(5.52)	(3.33)	(2.46)	(6.18)	(4.51)
Best rating		-0.00417			0.0290
		(-0.08)			(0.51)
R. Parker rating			-0.00829		0.0191
			(-0.26)		(1.23)
J. Robinson rating				-0.0192	-0.0129
				(-1.68)	(-0.59)
N	4058	3495	2681	1191	531
r2	0.690	0.668	0.668	0.740	0.774
aic	-6059.6	-4928.5	-4044.0	-3902.5	-1850.3
bic	-5845.1	-4712.9	-3849.5	-3744.9	-1722.1

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage×expert level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include aging, vintage×appellation, official ranking, type (color), re-rating year, and expert fixed effects. Ratings are corrected to span the 0-100 scale (see Equation 15). All ratings are in log.

Table F.4: Re-rating difference (new rating minus en primeur rating) as a function of the difference with estimated quality (en primeur rating minus estimated quality and expert bias), of the en primeur rating by the same expert, and of "salient" and best en primeur ratings.

	(1)	(2)	(3)	(4)	(5)
Theoretical adjustment	0.216^{+}	0.226^{+}	0.219^{+}	0.300^{+}	0.240^{+}
	(21.04)	(11.49)	(10.12)	(16.63)	(5.55)
En primeur initial rating	-0.499^{+}	-0.489^{+}	-0.526^{+}	-0.477^{+}	-0.519^{+}
	(-57.74)	(-28.75)	(-29.43)	(-34.27)	(-13.38)
Best rating		0.00644			-0.0243
		(0.43)			(-0.82)
R. Parker rating			0.0306		$0.0506^{\#}$
			(1.93)		(2.49)
J. Robinson rating				-0.0335^{+}	-0.0247
				(-4.00)	(-1.68)
N	12739	10426	6958	5270	2149
r2	0.682	0.678	0.695	0.804	0.792
aic	80565.0	66717.5	44790.7	30234.6	12484.7
bic	80848.2	67000.4	45044.0	30477.7	12683.2

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage×expert level. Significance levels: $^{\#}p < 0.05$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All regressions include aging, vintage×appellation, official ranking, type (color), re-rating year, and expert fixed effects. Ratings are corrected to span the 0-100 scale (see Equation 15).

G Some Micro-Foundations for the Empirics on Bordeaux Wines

Here we mention a couple of simple models that would micro-found the reduced form regressions on prices and re-ratings. As such, these models introduce specific assumptions that are not necessary, but provide one possible rationale for each situation. These models are adaptations of a recent approach Card and DellaVigna (2017) used in a different environment.

G.1 Prices

A wine has an unobserved quality q that is a function of some fundamentals f and of an independent term ϕ:

$$q = f + \phi. \tag{23}$$

An expert observes the fundamentals and a noisy signal of the other term: $s^r = \phi + \epsilon^r$ with $\epsilon^r \sim \Phi(0, \sigma^r)$. The superscript r denotes the considered expert, because this expert plays a role below as a "reference" expert influencing demand. Given the observed signal, the expert rates the item as

$$g^r = E(q|\, s^r, f) = f + E(\phi|\, s) = f + s^r, \tag{24}$$

with $E(q|\, s^r, f)$ denoting the expected quality conditioned on the observed s^r and f. In our application, this would be a typical "en primeur" rating of a Bordeaux wine, which most of the time isn't blind. Note that we do not consider the bias here to keep the notation uncluttered, but introducing it would be straightforward (just add it into the rating above).

Consumers are unbiased and can also observe the fundamentals. If the consumers aggregate a set of noisy and independent signals $s \in S$ that provide information about the term ϕ, then we can capture their expectation as $E(q|\, f, S)$.

Regardless of how many ratings a consumer observes, because of the salience of some particular expert's rating, the consumer could also be influenced directly by that rating. The consumer may also influenced by other factors such as the information printed on the bottle, e.g. the brand, the appellation and the official ranking. A simple way of thinking of this problem is to mix these factors, so that with some weight or probability λ the consumers base their expectation on a set of observed reviews S, with weight or probability μ they follow the signal on quality contained in the public information (the brand, appellation or official ranking) a, and with the remaining weight or probability $(1-\lambda-\mu)$, they follow the salient

expert's rating. The conditional expected quality or random consumer is then given by

$$\begin{aligned} E(q|\, g,f,S) &= \lambda E(q|\, f,S)+\mu a+(1-\lambda-\mu)\left(E(q|\, s^r,f)\right) \\ &= \lambda\widehat{q}+\mu a+(1-\lambda-\mu)g^r \end{aligned} \tag{25}$$

where $\widehat{q}$ is the best estimate of q given S (e.g., as the one we developed here).

As in the Bordeaux wine industry, quantities are essentially fixed for a given vintage, the main adjustment to increased demand is via prices. We thus estimate an hedonic regression of the form: $g_\theta(p) = E(q|\, g^r,f,S,b^r)$, where $g_\theta^{-1}(\cdot)$ is an increasing function that gives a price to a "perceived" quality in the market. In practice, we use the following version:

$$p=\beta\widehat{q}+\beta^r g^r+\nu_a+\nu_f+\nu_t+\nu_{sto}+\varepsilon, \tag{26}$$

where $g_\theta(\cdot)$ is assumed to be linear with slope θ, and with $\beta=\lambda\theta$, $\beta^r=(1-\lambda-\mu)\theta$. The other terms of the right hand side of Equation (26) control for effects found in the literature so far. The term ν_a denotes the official ranking fixed effect. We add a fundamentals fixed effect ν_f because it is likely that the fundamentals are not perfectly observed by the expert and could influence the price. The two other fixed effects, ν_t and ν_{sto}, capture the selling year and the retail store specifics that may also affect the posted price. ε is an error term.

The coefficients β and β^r are parameters of interest. We conjecture that our measure of true quality impacts prices, and so even when controlling for all determinants including for some salient experts ratings, β should remain positive and significant. Some of the previous literature suggests coefficient β^r may also be positive and significant.

G.2 Re-ratings

Next, consider a situation in which an expert, who already rated a wine/vintage "en primeur", re-rates that same wine. The expert observes two signals, s in the first period (en primeur), as well as a new conditionally independent signal s', so that $s=\phi+\epsilon$ and $s'=\phi+\epsilon'$ with $\epsilon,\epsilon'\sim\Phi(0,\sigma)$ and $\epsilon'\perp\epsilon$. In the first period, every thing works as before, that is as in Equation 24 (dropping r superscripts). In the second period, the expert's rating may be dependent upon her own previous signal. Moreover, the expert could be also influenced by peers, and in particular by the most prominent ones. Therefore the expert's re-estimation of quality is $E(q|\, s,s',s^r,f)$, which is conditioned on the fundamentals f, the previous signal s, the new signal s', and the "reference expert" rating g^r (which, for instance, leads the expert to know the other prominent expert's signal s^r). The new rating g' is thus given by:

$$g'=E(q|\, s,s',s^r,f)=f+E(\phi|\, s,s',s^r). \tag{27}$$

Again, as a simplifying assumption, suppose that the expert weights the first signal with prob λ, the new signal with prob μ, and the reference expert signal with prob $(1-\lambda-\mu)$. Equation (27) becomes

$$g' = f + \lambda E(\phi|\, s) + \mu E(\phi|\, s') + (1-\lambda-\mu)E(\phi|\, s^r).$$

Using Equations 23 and 24, this becomes:

$$g' = f + \lambda g + \mu\left(\widehat{q} - f + \epsilon'\right) + (1-\lambda-\mu)g^r.$$

Rearranging and adding fixed effects and error term, we get the following equation:

$$g' = \beta_1 \widehat{q} + \beta_2 g + \beta_3 g^r + \nu_a + \nu_f + \nu_t + \nu_e + \epsilon', \tag{28}$$

where $\beta_1 = \mu$, $\beta_2 = \lambda$ and $\beta_3 = (1-\lambda-\mu)$. As before, ν_a denotes official ranking fixed effects and ν_f a vintage/appellation fixed effect that captures the fundamentals. The term ν_t accounts for the re-rating year and ν_e is an expert fixed effect. ϵ' is the error term.