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

NICOLAS CARAYOL AND MATTHEW O. JACKSON *

Draft: January 2021

Abstract

Consumers and businesses 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 processes ratings by a group of reviewers over a set of items and simultaneously evaluates the individual reviewers' accuracies and biases, in a way that provides 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 ratings. 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 experts and 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: GRETHA, Department of Economics, 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. 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 seminar participants in Paris, Melbourne, and Bordeaux. 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 evaluated and rated. Prominent examples include films, theater, art, books, wines, restaurants, stocks, and most consumer products. The internet and various platforms have led to even enormous growth in the number of items that are evaluated and the number of people doing the rating. Selling platforms on the web most often report previous consumers ratings. Numerous other websites collect evaluations from distributed sources and report them to the public. These ratings can come from experts (movies critique ratings) or from users (e.g. Yelp). In fact, an important innovation that has accompanied the "digital economy" is that ratings are now nearly freely available about almost everything.

As market efficiency improves when products' and services' qualities are better assessed (Akerlof, 1970), platforms can gain from the availability and use of ratings, exploiting the so-called "wisdom of the crowd" famously illustrated by Galton (1907). Under suitable conditions, collecting a large number of independent individual evaluations and computing the mean (or median) provides a reliable estimate of the truth. Though there is substantial noise in the ratings as individuals can have large biases and vary widely in their accuracy, positive and negative errors cancel out when data points become sufficiently numerous.

Those conditions, however, are rarely met even in the internet age. For example, consider the ratings data of the largest and widest such information aggregator over more than twenty years, *Amazon.com*, out of the 182 million verified ratings in this database, 80% of the 12.1 millions rated products have fewer than 9 ratings, and 90% of all products have less than 22 ratings. In fact, out of the 21 product categories at *Amazon.com*, 19 were such that more than two-thirds of the items had no more than 10 ratings.[1] Therefore, simply averaging ratings likely provides biased and noisy estimates for the vast majority of items, given how few ratings they have. Often, consumers are reluctant to buy products which have received few ratings, and inefficient rational herding may occur (Banerjee, 1992; Bikhchandani, Hirshleifer and Welch, 1992).

An average under-weights the ratings of people who are very careful and discerning and over-weights others who are careless and frivolous. An average is also susceptible to deliberate manipulation and to bias. For instance, some people only rate items with which they have extreme experiences, leading to a selection bias where they post excessively extreme ratings.[2]

An optimal system should instead take into account a person's full set of ratings *across items* to evaluate their biases and accuracies. By undoing biases and putting more weight on the most accurate reviewers, well-processed ratings can result in large improvements compared to an average - especially for the multitude of items that are rated by small numbers of people. 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

[1] For details on Amazon data, see Table 5 and Figure 5 in the Online Appendix A.

[2] For instance, see the discussion in Nei (2017).

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. Thus, to be fully effective, a system for processing ratings must simultaneously estimate three things: 1/ the 'true qualities' of the items, 2/ the accuracies of the people who are rating the items, and 3/ any bias that each particular reviewer has.

In this paper, we develop a model and provide an estimation technique to do these three things simultaneously in a consistent manner. Our identification makes use of the ratings of a set of people over multiple items: a person's ratings on other items are used to discern their bias and accuracy and then used to adjust and appropriately weight their rating on any given item. Specifically, we show how this can be done via a slight variation of well-known weighted regression methods that simultaneously estimate items' true qualities and individual raters' biases and accuracies.

While we refer to 'true qualities' throughout, we emphasize that tastes are subjective. What we mean by 'true quality' is an 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 strong predictors of what many people's ratings would eventually converge to - after adjusting for each person's systematic tendency to be more favorable (or less favorable) than others on all items. More accurate reviewers' ratings are thus more valuable in predicting the (subjective) consensus. A reviewer with a low accuracy might still have "good taste" in some other sense, but would not be as useful in predicting the consensus rating.

Our method provides some immunity to well-known basic types of manipulation of ratings, as well as selection biases in ratings.[3] If a rater only provides high ratings, then that bias can be identified. If a reviewer only provides 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, which thus counters some important forms of manipulation and systematic bias.[4]

Note our analysis is very different from simply adjusting different reviewers' scales. Such an adjustment is part of the process, but only one part. To understand this first part of the adjustment, note that even when reviewers use the same scale, they may have different distributions of ratings. For instance, one reviewer may use 100 point scale and only use

[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 highly accurate rater who has an impeccable history of accurate ratings tries to manipulate a single item's rating in a case when there are very few other ratings, then one could never be sure that there was manipulation. But if someone simply provides high ratings to some subset of items, then that is easily identified and eliminated. Also, our system takes into account the history of a reviewer, and so the ratings of shills will largely be ignored.

the range between 70 and 100, with an average of 87, but rarely ever give a score below 70. Another reviewer may only give scores between 50 and 100 with an average of 80. Another reviewer may use a scale from 1 to 20 rather than 0 to 100. The fact that this can impact some aggregate of reviewers' scores 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. One can normalize scores by aligning some order statistics of the distributions and then also translating them to a common scale, so that the variances are the same for each reviewer. A 90 from one reviewer might be translated to an 88 from another reviewer.[5]

Having normalized the scales so that they are comparable, one approach is to average reviewers' ratings at that point.[6]

The improvement in estimation comes after this scale adjustment – from accounting for the facts reviewers still have biases, and some reviewers are more accurate in matching the true value than others. Assessing the true value then is a statistical exercise that involves combining the ratings in a way beyond just averaging them, but by estimating the true values that are most likely to have produced the observed ratings. Intuitively, an estimate of a true quality adjusts for reviewers' biases and weights different reviewers' ratings to adjust for their overall accuracy. Those implicit weights and bias adjustments depend on the set of reviewers rating any particular item, and adjust with the overall set of items being rated in a way that minimizes the overall errors in the system.

We demonstrate that our estimates of item quality, evaluators biases and accuracies are unbiased and consistent. Using Monte Carlo simulations, we further show that our method provides more accurate estimates of true underlying qualities than average ratings, and by substantial amounts even with more ratings than typically appear in online systems. This happens not only in small data sets, but also in larger ones. In fact, the accuracy improvement can be a third or even more, and it increases as the number of items being rated increases. Having more items enables us to more accurately estimate evaluators' biases and accuracies, which in turn improves our estimates of the qualities.

After presenting our method and its properties, we illustrate it on '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 have an interesting relationship (Dubois

[5] Ashenfelter and Quandt (1999) instead convert reviewers' scores into a ranking. That works when all reviewers are reviewing exactly the same set of objects, but does not work more generally. The renormalization that we use can be thought of a generalization of such a ranking method, since it preserves each reviewers rankings and puts them on a common scale, but allows them to be rating different sets of items.

[6] For instance, see the description of how the "Global Wine Score" is constructed (https://medium.com/the-global-wine-score/what-are-the-differences-between-the-global-wine-score-and-users-ratings-systems-5ae0fd64926e , accessed October 16, 2017). See also the discussion in Ekstrand, Riedl and Konstan (2011).

and Nauges, 2010; Friberg and Gronqvist, 2012; Hilger, Rafert and Villas-Boas, 2011) 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 and discuss their individual biases and accuracies, and v) experts rate quasi-simultaneously every year and independently all wine vintages before bottling at the "en primeur" stage, thereby minimizing inter-expert influence. Our empirical results are not only of interest to those specifically interested in wine markets, but also as an example of informational issues that occur in many different markets.

This dataset includes a very comprehensive set of wines, as well as the raters identity and posted prices in retail outlets in three major markets (Paris, New York and Hong Kong). We use our quality estimates derived from these data 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 fourteen percent increase in price: so an elasticity of 1.4. In particular, our index remains significant when also accounting for prominent expert ratings (Parker and Robinson); as well as the highest ratings, since many retail outlets selectively quote the highest ratings for each wine. We also find that that the elasticity of a wine price in an expert's ratings increases with our estimate of that expert's accuracy. In an Online Appendix, we also use re-rating data – experts often re-taste and re-evaluate the exact same vintage of the same wines at later dates – and show that the adjustment in ratings is strongly predicted by our estimated quality, controlling for many other factors and fixed effects.

The idea of combining a number of opinions and judgments is by no means novel. Aggregating opinions has been discussed since Condorcet (1785), and a topic of importance following Arrow (1951), and spans from individuals processing information from multiple sources (e.g., see Budescu, 2005) to the literature on herding (e.g., see Banerjee, 1992; Bikhchandani, Hirshleifer and Welch, 1992). Our results also connect with the literature on the impact of ratings on markets (e.g. Ginsburgh and van Ours, 2003; Reinstein and Snyder, 2005; Finkelstein, 2009; Luca and Smith, 2013; Luca, 2016; Dai et al., 2018), including reviewers' incentives (e.g., Nei (2017); Dai et al. (2018)). However, those studies do not aim to improve quality estimates from existing information. There have been previous studies, such as that by Budescu and Chen (2015), which 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 accordingly 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 from the formulation of the problem, and the adjustments in identification that then allows us to estimate the qualities, biases, and accuracies in a consistent and unbiased manner. We examine situations in which raters have possibly different and overlapping histories of rating previous items, and our technique *simultaneously* estimates true qualities of the items together with the biases and accuracies of the raters.

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. By providing a new technique that improves the quality of item ratings, and simultaneously uncovers how accurate reviewers are, our analysis complements the previous literature on recommender systems.

The paper is organized as follows. In Section 2, we introduce the setting, and in Section 3, we present our approach, and in Section 3.4, we illustrate the ideas behind our approach on a famous case study, namely the Paris 1976 wine contest. In Section 4, we demonstrate the properties and added value of our estimation both analytically and via Monte Carlo simulations. In Section 5, we apply the approach on a new and large 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 further demonstrate the predictive power of the approach via a study of experts' adjustments of ratings of wines they earlier rated as "en primeur".

2 Evaluators, Items and Ratings

2.1 Notation

A set N of items $i = 1,...,n$ is to be rated.

A set M of evaluators $j = 1,...,m$ each rate a specific subset of the items $N_j \subset N$.

We use the term evaluators as a generic term encompassing users, reviewers, critics, and experts. Evaluators 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 evaluator rates an item at most once, and so 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 information) indicating that j did not rate item i.

Let 1_{ij} be the indicator variable that is 1 if evaluator j rated item i, and 0 otherwise (so it is the indicator that $g_{ij} \neq .$). 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 evaluator j. The total number of ratings is given by $R = \sum_{ij} 1_{ij}$.

2.2 Evaluators' Ratings of Items

When rating an item i, an evaluator j (independently) estimates the unobserved true quality of that item q_i. Evaluator j may have a systematic bias b_j (for instance always over- or always under-rating items). Moreover, evaluators are not perfect evaluators and so each rating is likely to include an error ε_{ij} on the top of the systematic evaluator-specific bias. A simple and natural way to take those three dimensions (true quality, bias and error) into account is to let evaluator i's observed rating be defined by

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

where the errors ε_{ij} are centered, independent across js and i.i.d. for the same j, with standard error σ_j.

A measure of the "*accuracy*" of evaluator 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 do rate. For instance, it is ok if a wine evaluator only chooses to rate wines that sell large quantities, or small quantities, or have received high past reviews, 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). From seeing enough observations of ratings, regardless of how they were selected – provided they have independent error terms – we can estimate the accuracy and bias.

2.3 "Quality" and Subjective Tastes

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 just the average rating if an infinite number of unbiased people all rated this item (perhaps subjectively). This is still a very useful exercise since people have correlated tastes and knowing this answer can help people predict how much they will personally enjoy the item if they consume it. Thus, even though we use the term "true quality," this should be interpreted as a "true anchor," around which people may disagree, but then when they disagree it comes from two sources: a systematic idiosyncratic bias in their taste and a term which is essentially random from our perspective.

In this light, the appropriate interpretation of 'bias' is divergence from that anchor. For instance, if we examine the ratings of a set of romantic comedy movies, it might be that some particular rater tends to like this genre more than other raters, providing relatively inflated ratings, and thus has a positive bias. After seeing a person's ratings on a few romantic comedies we can begin to estimate their bias, and with many instances we can accurately estimate their bias. Their ratings are still useful since once we adjust for that bias, their *relative ratings* of the different movies provides valuable information. Even raters for whom we have only a few ratings are useful, as they are still providing information.

3 Estimating Items' Qualities and evaluators' 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

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

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 evaluator j, and ε_t has variance σ_j^2 where j is the evaluator 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:

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

The weighting is by the precision or accuracy of each evaluator, 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 an evaluator 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 difficulty is that there are interdependencies in the I_{it} and I_{jt}. In particular each observation r corresponds to one item i and one evaluator j, and no two observations have the same combination of items and evaluators.

To see this in more detail, 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 evaluators 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 evaluators' biases by the same amount, then we still have a solution to these equations.[7]

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 evaluators that estimate the same pair of items, the sum of the rows corresponding to evaluator 1 on item 1 and evaluator 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 evaluator 1 on item 2 and evaluator 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. For instance, it would be strange to say that *all* reviewers over-estimate qualities. 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 rank of the matrix. 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 evaluators $M = M_1 \cup M_2$, such that evaluators in M_1 only rate items in N_1 and evaluators 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 raters in M_1 and correspondingly decrease the value of the items in N_1; and then do the reverse for the evaluators in M_2 and items in N_2, in a way that 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 evaluators - every item is rated by at least one evaluator, and we can find a cycle defined by overlaps that includes all evaluators such that evaluator 1 overlaps in the set of items with evaluator 2 who overlaps with evaluator 3, and so forth.

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 evaluator.

[7] 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.

3.3 A Two-Stage Estimation

Once we have the appropriate normalization, and sufficient rank of the item-evaluators matrix we then run a heteroscedastic-consistent two-stage weighted least squares analysis.[8]

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

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

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

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 evaluator on observation ij (or more generally, for the observation r if there are multiple ratings of the same item per evaluator).

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

$$q_i^{one} = \sum_j \frac{1_{ij} g_{ij}}{m_i}, \tag{9}$$

and we can estimate the b_j's via

$$b_j^{one} = \sum_i \frac{1_{ij} \left(g_{ij} - q_i^{one}\right)}{n_j}. \tag{10}$$

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

$$(\sigma_j^{one})^2 = \sum_i \frac{1_{ij} \left(g_{ij} - b_j^{one} - q_i^{one}\right)^2}{n_j}. \tag{11}$$

However, more generally, if some evaluators 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 evaluators, and so their biases all cancel out, greatly simplifying the estimation. When not all evaluators evaluate all items, then these no longer cancel and one has to solve the standard regression equations for the

[8] 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.

normalized matrix.

We remark that one cannot iterate any further on this procedure. In particular, if we iteratively re-estimate the q_is with the new weights based on new estimates of the errors σ_js from the second stage, then the process can converge to setting some $\sigma_j = 0$ and the $q_i = g_{ij}$, which becomes self-fulfilling.

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 1976 wine tasting that included red wines from both California and Bordeaux, since this had important consequences and since others, including Lindley (2006) and Ashenfelter and Quandt (1999),[9] have used it to discuss methods of recombining experts' ratings. 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. Let

$$q_i^{avg} = \sum\nolimits_j \frac{1_{ij} g_{ij}}{m_i}. \tag{12}$$

The variance in the experts' ratings, as noted by Lindley (2006) for instance, has generated some 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 raters, 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}, 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 does result 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. Our estimates of experts' biases and accuracies exposed in the bottom part

[9] We include only the red wines, as the data on the white wines have some issues, as discussed in Hulkower (2009).

of Table 1 show that whereas biases relatively small, there is a large variation in experts' accuracies.

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.

Wine and Vintage	Our Quality Estimate q_i^{two}	Average Rating q_i^{avg}	Borda Score 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)	6.13	0.90	-0.84
Christian Vanneque (sommelier, restaurant La Tour D'Argent)	17.40	0.32	0.11
Claude Dubois-Millot (sales director, guide Gault et Millaud)	5.22	1.06	-0.24
Jean-Claude Vrinat (owner, restaurant Taillevent)	3.29	1.68	-0.14
Michel Dovaz (Institut du Vin)	6.15	0.90	-0.29
Odette Kahn (director, La Revue du Vin de France)	13.05	0.42	-2.64
Patricia Gallagher (l'Académie du Vin)	6.24	0.89	2.06
Pierre Brejoux (inspector general, Institut des Appelations d'Origine)	8.42	0.66	0.16
Pierre Tari (owner, Chateau Giscours)	5.82	0.95	1.66
Raymond Oliver (owner and Chef, Restaurant Le Grand Vefour)	2.78	1.99	-0.24
Steven Spurrier (l'Académie du Vin)	4.52	1.22	0.36

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

It is interesting 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). However, those latter three experts 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 more tempered but more reliable 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.

This example shows how our method can differ from both the average rating and a Borda ranking. Since we have no known "truth" with which to compare these rankings the example just provides some insight into the types of adjustments that our approach makes. Thus, to provide more insight about how our estimates compare to the truth, we next explore the properties of our estimates, and compare them to the average, via analytic results and simulations.

4 Properties and Gains of the Two Stage Estimation

In this section we study the properties of estimates q_i^{two}, b_j^{two} and σ_j^{two}. We do this via analytic results about our estimators complemented by 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 evaluators' 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 raters has support $\left[\underline{\sigma}_\varepsilon^2, \overline{\sigma}_\varepsilon^2\right]$.[10]

Let each evaluator review at least $n(R)$ items and each item be reviewed by at least $m(R)$ evaluators, 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}, b_j^{two}, are consistent (e.g., see White (1980)), and then given the growing $n(R)$, so are the estimated variance terms

$$\left(\sigma_j^{two}\right)^2 = \frac{\sum_i 1_{ij} \left(g_{ij} - b_j^{two} - q_i^{two}\right)^2}{n_j}. \tag{13}$$

We also provide conditions under which the estimates are unbiased.

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

LEMMA 1 *In addition, if the evaluators' 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.*

The 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

$$E\left[(q_i^{avg}-q_i)^2\right]=\frac{\sigma_b^2}{m(R)}+\frac{\sigma_\varepsilon^2}{m(R)}. \tag{14}$$

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

When we instead use our approach, then the biases estimated and 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 get a better sense for how well our method estimates true qualities and how it compares to other methods, we perform Monte Carlo simulations in which we can know what the true qualities actually are and then directly measure how different methods compare.

Item qualities are randomly drawn from a uniform distribution $q_i \sim \mathrm{U}\left(\underline{q},\overline{q}\right)$. Evaluators' 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 quality and evaluators' biases and accuracies, ratings are generated according to Equation (1). Since in many applications, not all evaluators 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 evaluators 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 evaluators' 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).[11]

Figure 1 plots our estimates of the item qualities, the biases, and the average variances – each against the corresponding true values. The left graphs are based abstract values of the parameters whereas right graphs correspond to calibrated data on the Paris 1976 wine contest. We can see that all points are very close to the 45 degree line indicating that for those numerical conditions, the two stage procedure is providing a pretty good estimation

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

of items' qualities and evaluators' biases and accuracies. Next, we explicitly measure the performance of our estimates.

Next we investigate how well our method compares to the average rating. The comparison measure, which we call *Gain*, is defined as the extra share of the per item quality that is explained by our method compared to the average ratings

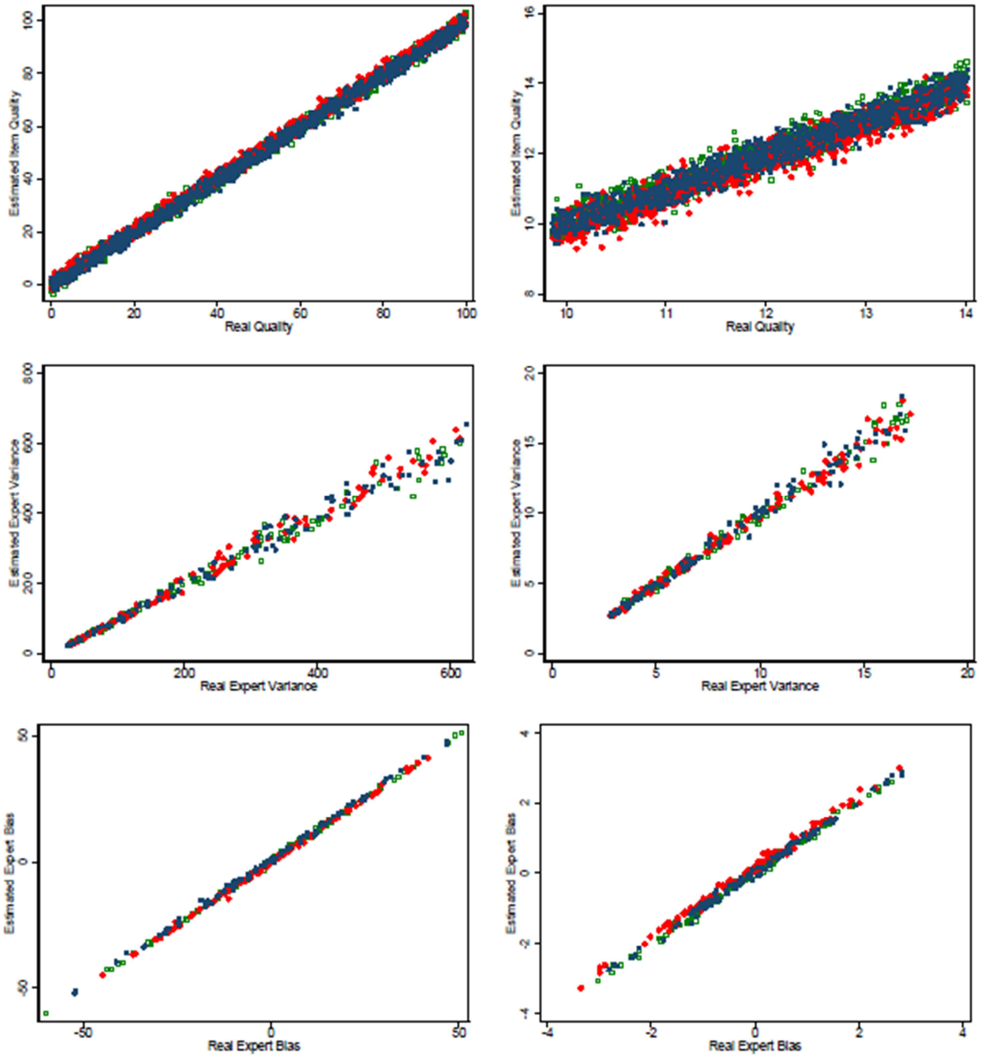
$$\text{Gain} = 1 - \frac{\sum_i \frac{\left(q_i^{two} - q_i\right)^2}{n}}{\sum_i \frac{\left(q_i^{avg} - q_i\right)^2}{n}}. \tag{15}$$

Figure 2 presents this measure for a series of Monte Carlo simulations, varying parameters. In the left graph, we vary both the number of items and the number of evaluators. We see that gain increases in both the number of evaluators and number of items. This reflects two forces: more items enables us to learn more about each evaluator and better estimate their bias and accuracy, and so better estimate qualities, while more evaluators gives us better estimates of the qualities just through having more draws of scores. The gain relative to average scores is substantial and increases with the number of items and evaluators, approaching 50 percent. In the right graph, we keep the number of evaluators fixed and vary biases and accuracies across different numbers of items. The main insight is that evaluator accuracy is important in resolving noise in the data. It is not necessary that all evaluators be accurate, which is why increasing the homogeneity in evaluators' accuracies lowers model fitness and gain more than does decreasing their accuracy on average. The model performs better when evaluators are less biased, but this is due to the fact that they have rated only parts of the items here. In the Online Appendix B.1, Figure 6 presents the same graph when evaluators have rated all items. Evaluators' biases are then more precisely estimated, which leads to increased performance of the estimation and gain over the average rating when evaluators are more biased.

In the same Online Appendix, Figure 7, we compare different random assignments of which evaluators rate which items with the same number of edges (rating observations). When the ratings matrix is sparser (evaluators 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).

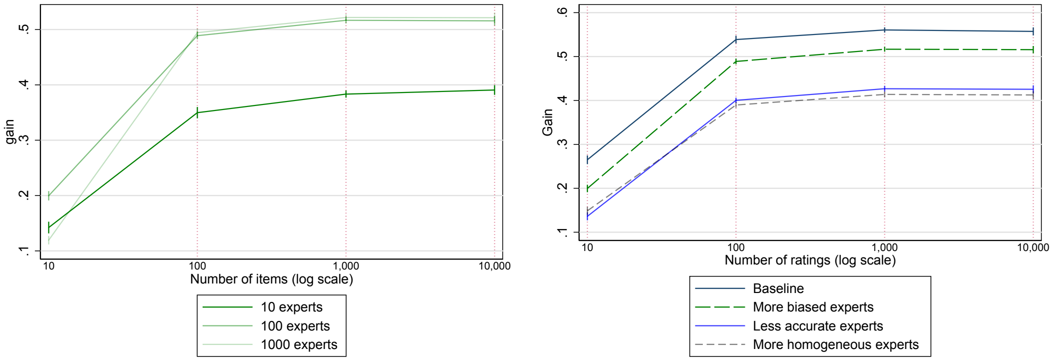
We also compare our estimates to using a Borda score (see above). The Borda score is not intended to provide an estimate of quality (an index), but a ranking. We thus transform real quality and estimated quality into ranks, and then compare them to the Borda rankings. The standard measure of rank correlation 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 with only 10 items; and Borda beats

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 $\overline{\sigma} = 4.17$.

Figure 2: Monte Carlo numerical experiments.



Notes: Fractional polynomial estimates and 95% confidence intervals where each graph point corresponds to 1,000 Monte Carlo data simulations. In the left graph, $f = 0.5$, $\underline{q} = 0, \overline{q} = 100, \sigma_b = 10, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. In the right 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 "More biased evaluators" assumes $\sigma_b = 20$. In the "Less accurate evaluators" case, we assume $\underline{\sigma} = 10$ and $\overline{\sigma} = 30$, whereas in the "More homogeneous evaluators" case, $\underline{\sigma} = 10$, and $\overline{\sigma} = 20$.

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 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	.54	.41	.47	.49	.51	.40
100	.88	.12	.14	.86	.95	.05
1,000	1	0	.01	.99	1	0
10,000	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 presumably very large (e.g., prices vary significantly) and difficult to appreciate (as 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). Official rankings and expertise have historically played a very important role in the development of these markets (Powell, 2017). However, experts' opinions have been shown to diverge even within relatively homogeneous sub-segments of the market (Ashton, 2012; Hodgson and Cao, 2014). Some authors even expressed doubts about the added information contained in those ratings (Ashenfelter, Ashmore and Lalonde, 1995).

We now 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 even bottled are tasted and rated by trained professionals and experts. Their ratings are vital for intermediaries and investors who will buy most of the production. Many of these ratings are eventually published in various media (magazines, books, websites). The wine will only be 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.

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 critiques. In some cases an expert has issued multiple ratings of the same wine and vintage. In those cases we use the mean rating. This results in $51,363$ ratings.[12] We also 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 in Online Appendix C presents the distribution of wives and ratings across vintage years.

[12] The analysis is robust to dropping the bottom five percent of the wines.

5.1.1 Scaling the Ratings

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.[13] 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.

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

$$g_{ij} = 100 \times \left(G_{ij} - G_j^L\right) / \left(G_j^H - G_j^L\right), \tag{16}$$

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

Figures 9 and 10 in Online 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.[14]

5.2 Estimating Experts' Biases and Accuracies

In Table 3 we summarize experts' estimated characteristics. Figures 11 and 12 in Online Appendix C 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

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

[13] The lower tail of ratings is quite long and noisy, and so in Online Appendix C we provide the same analysis when the lowest five percent of ratings are dropped. It makes little difference given the size of the data, but could be improving in some other settings where there are occasional peculiar items.

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

Therefore,

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

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.[15] We study the relationship between accuracy and correlation and find a positive relation between the two indicators, but they are clearly distinct (see Figure 13 in Online Appendix C).

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 C (Figure 16). We see little relationship between errors and quality from the fifth percentile of item quality.

5.3 Estimating Wine Qualities

We present the top-100 wines from the sample along with their estimated qualities in Table 8 in Online Appendix C. The number one Bordeaux wine is actually a Sauterne (sweet white wine), Chateau Yquem 2009, and Chateau Marguaux 2010 is the best red wine.[16]

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 would probably sounds weird 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 14 in the Online Appendix C 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 8 in Online Appendix C.

[15] 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.

[16] 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.

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	1.00	0.79	1.45	1,140
Bettane et Desseauve	64.46	1.12	0.82	7.98	3,011
Chris Kissack	88.19	0.82	0.79	0.90	2,431
Decanter	65.11	1.11	0.88	-9.83	2,342
Jacques Dupont	149.77	0.48	0.69	-18.32	3,077
Jacques Perrin	91.71	0.79	0.89	-22.13	488
James Suckling	81.56	0.89	0.82	-3.30	1,985
Jancis Robinson	80.59	0.90	0.69	-1.26	3,793
Jean-Marc Quarin	39.62	1.83	0.87	0.57	3,042
Jeannie Cho Lee	51.00	1.42	0.84	14.42	1,308
Jeff Leve	67.32	1.08	0.89	-3.07	1,530
La Revue du Vin de France	91.72	0.79	0.80	-1.63	2,216
Neal Martin	52.22	1.39	0.82	12.01	2,965
Rene Gabriel	70.82	1.02	0.80	8.96	4,757
Robert Parker	56.78	1.28	0.81	13.14	2,838
Tim Atkin	73.69	0.98	0.75	7.51	1,900
Wine Enthusiast	118.89	0.61	0.75	-5.10	2513
Wine Spectator	74.84	0.97	0.84	5.45	3,669
Yves Beck	141.53	0.51	0.78	-7.74	441

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 C.2 (Table 9 and Figure 15). 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

10 and 11 and Figure 17.

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. 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": 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:

$$g_{ij} = q_i + b_{j,k} + \varepsilon_{i,j,k}, \quad \forall i \in k, \forall k \in \{L, N \backslash L\} \tag{17}$$

Thus, experts have category-specific biases that are interesting to compare. The differences in the estimated biases across the left vs right dichotomy are:

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

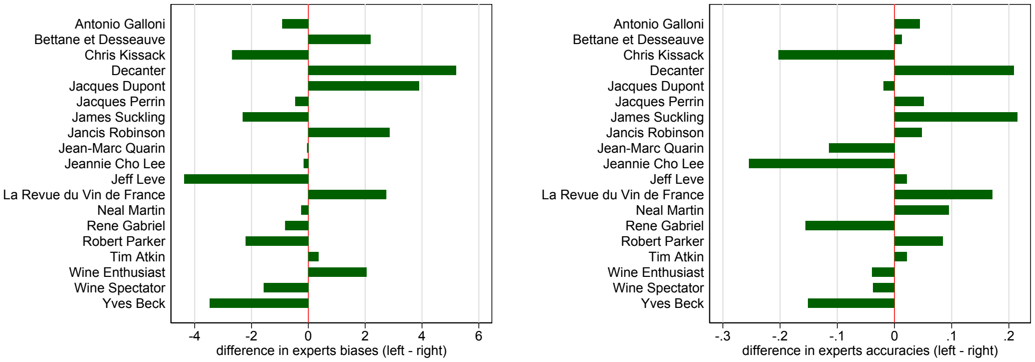
and similarly differences in expert j's normalized accuracies between Left and Right Bank wines as

$$\Delta A_j^{two} = A_{j,L}^{two} - A_{j,N\setminus L}^{two}, \tag{19}$$

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 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. This could explain partly the well known Pavie 2003 controversy[17] 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).

There is, however, no clear correlation pattern between the differences in accuracies and the differences in biases (and see Figure 18 in Online Appendix C 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.

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.

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

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:

$$RSS_1 = \sum\nolimits_{i,j} 1_{ij} \left(g_{ij} - b_j^{two} - q_i^{two} \right)^2 A_j^{two}. \tag{20}$$

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

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

this becomes

$$RSS_1 = \frac{R}{m} \sum_{j'} \left(\sigma_{j'}^{two} \right)^2 \tag{21}$$

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

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

Using the similar calculations as for Equation 21, it comes:

$$RSS_2 = \frac{R_L}{m} \sum_{j'} \left(\sigma_{j',L}^{two} \right)^2 + \frac{R_{N\setminus L}}{m} \sum_{j'} \left(\sigma_{j',N\setminus L}^{two} \right)^2, \tag{22}$$

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 have 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 20.32 greatly exceeds that threshold value. Thus, there are significant differences in experts' rating patterns for Left and Right Bank wines.

6 Bordeaux Wines Estimated Qualities, Experts' Accuracies and Prices

In this section, we use the expected positive relation of wine quality on prices as well as an expected higher correlation of accurate experts ratings with wine prices, to assess whether our methodology captures unobserved information on Bordeaux wines.

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 are very much linked to the notion of "terroir" as they relate to specific sub-regions of production as well as (most of the time) typical 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 6 in Online Appendix C). Besides, we know when wines are "first wines" of a "chateau" listed in one of the official ranking of the Bordeaux production area, such as Grand Cru Classe 1855 or Premier Grand Cru Classe A (see Table 7 in Online Appendix C?).

Prices and Store/Market Data The prices of the wines are from surveys of restaurants in three of the main world-wide markets: in Hong Kong, New York and Paris. In these cities, the wine prices of, respectively, 244, 437 and 409 restaurants were surveyed at different points in time (some more details in Table 12, Online Appendix D).) 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 prices observations, in a given shop and year. Out of the 2,871 wine/vintage that we consider, we have 43,307 such observations, that is 15.08 prices on average for each wine/vintage.

In Online Appendix D, Figure 19 shows the price distributions in the three markets and Table 13 lists the top-100 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 nearly fixed for a given vintage.[18] 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 are clearly observed by the consumers and are likely to 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.

Besides, Ashenfelter, Ashmore and Lalonde (1995) and Ashenfelter (2008) highlight that wine prices are affected by the weather conditions at crucial points in the season in the production year and by wine aging. We control for such weather conditions by including vintage-appellation fixed effects: dummies that capture the weather conditions for various vintages in the specific sub-region of Bordeaux production.

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

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 "reference" experts' ratings by directly including the ratings

[18] Production cannot be significantly adjusted upward by mixing the wine of a vintage with wine from other vintages: 85% of the wine must come from property grapes of the referenced vintage.

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. Lastly, as a limit experiment, we also use the average rating among experts (properly normalized) as a supplementary control to check whether our estimated quality still significantly explains price variation even controlling for the average rating. We can not go further as estimated quality, average rating, and other ratings are in general strongly correlated. All ratings used are corrected to span the 0-100 scale (as exposed in Equation 16).

Section F.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 significant predictor of prices. Its coefficient is positive and always significant. In the first five estimations, when a list includes our estimated quality, it is significant at the .001 level. In the a particularly telling regression (column 5), in which the ratings of famous experts, Robert Parker and Jancis Robinson, are included as regressors, our estimated quality significantly explains the price whereas other different ratings do not. Interestingly, Robert Parker who is often considered as a "guru" and is claimed to influence prices, ends up having no significant influence on prices after controlling for our estimated quality.

As prices and ratings are in logs, the coefficients can be interpreted as elasticities. The elasticity of prices on item quality is very high. According to the comprehensive regression estimates in column 5, a 10 percent increase in quality raises the price by nearly 35 percents. This is consistent with the idea that, in this industry, there is a high elasticity of prices with respect to quality.

Even when we introduce the average rating as a regressor (in column 6), the coefficient of our quality estimate remains significant (at the .05 level) and positive even though both are, of course, positively correlated. Thus, our quality estimates still significantly improves the prediction of prices after the average rating is taken into account. Note that this is even though these average ratings already incorporate an adjustment in which all experts' ratings are put on the same scale.

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

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)
Estimated quality	2.526^{+}	2.402^{+}	2.584^{+}	2.892^{+}	3.306^{+}	$0.737^{\#}$
	(17.98)	(10.42)	(11.89)	(17.24)	(9.57)	(1.90)
Best rating		0.172			-0.248	
		(0.87)			(-0.79)	
R. Parker rating			0.126		0.0954	
			(0.55)		(0.31)	
J. Robinson rating				$0.166^{\#}$	0.0754	
				(2.19)	(0.76)	
Average rating						1.893^{+}
						(5.38)
N	43307	43307	36109	28774	23974	43307
r2	0.801	0.801	0.798	0.821	0.828	0.803
aic	56447.8	56431.7	46489.6	32524.4	26316.4	55901.6
bic	59137.4	59138.7	48689.7	34549.9	28022.3	58599.9

Notes: t-statistics are in parentheses. The standard errors are clustered at the wine×vintage level. Significance levels: $^{\#}p < 0.1$, $^{\star}p < 0.01$, $^{+}p < 0.001$. All variables (dependent and explaining) are in logs so that coefficients can be interpreted as elasticities. All regressions include vintage×appellation, official ranking, year, type (color), and store fixed effects. All ratings are corrected to span the 0-100 scale (see Equation 16).

Bordeaux wine quality. It is also providing external support to our item quality estimation methodology. We would now like to assess another important output of our methodology: an estimation of experts accuracies. Are estimated expert accuracies also consistent with price data?

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 expurgate the correlation between prices and experts' ratings from external confounding factors, we first regress log prices on each expert's logs ratings separately, controlling for a number of factors such as the rating year, the interaction of vintage and appellation dummies (which captures in particular local weather conditions in the production year), official ranking dummies, wine type (red, white or sweet) dummies, and retail shops fixed effects. Raw regression results are exposed in Table 14, Online Appendix D). As both ratings and prices are in logs, we can interpret those coefficients as, for each expert taken separately, her ratings-elasticity of wine prices.

In a second stage, we study the relation between those estimated elasticities and experts accuracies. Among the thirteen experts considered,[19] 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 accuracy against their ratings-elasticity of prices. 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 strongly support the idea that the correlation between an expert's ratings and prices increases with estimated expert's accuracy. Some experts do lie above or below the confidence interval though. Bettane et Dessauve and Robert Parker have a correlation with prices that goes beyond what is predicted by their accuracy. Some others, namely 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.[20] It could also 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).

Figures 15–18 in Online Appendix D 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

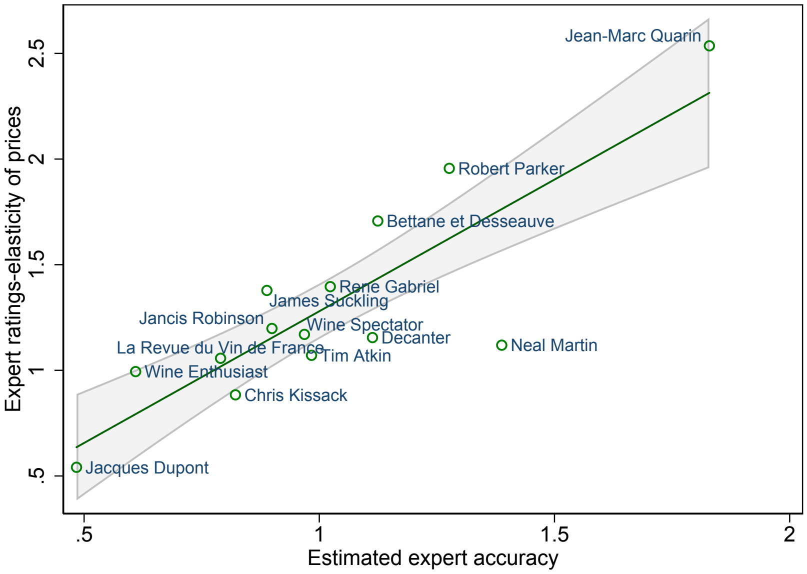
In Online Appendix E, we examine another aspect of our Bordeaux wine quality estimations relying on a very different design, making use of additional rating data. It turns out that some experts rate the exact same Bordeaux wines at (at least) two different points in time: a first time "en primeur" (those 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.

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. There is no reason a priori that a second rating comes closer to our estimated quality except that there is a regression to the real quality. Results remain when we consider that experts may also be directly influenced by other experts' opinions, and therefore control, as in the

[19] 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).

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

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



The vertical axis lists correlations of ratings with prices controlling for vintage, vintage×appellation, official ranking and retail shop fixed effects.

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.

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 3.5) predictor of wine prices, even when controlling for many fixed effects and other measures of ratings, including an average rating (which is insignificant). We also show that our estimate significantly predicts experts' re-ratings, so that their second ratings move closer to our estimated quality over time (while, again, an average rating is not a significant predictor).

The fact that our technique not only identifies estimates of item true qualities, but also provide estimates of raters' biases and qualities should also be very valuable to various systems. For instance, one can provide incentives for most accurate raters to provide ratings on particular items that may be in high demand. With our estimates of raters' accuracies, we know how much improvement in the accuracy of an item's estimated quality we will get by having that rater provide a rating. 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. For instance, some expert may be more accurate and less biased in rating Bordeaux than Rhone wines.

We close with a note on a further extension and application of our techniques.

The analysis above presumes that reviewers may have some bias, but then their reviews are then randomly distributed around the true quality subject to that bias. That is, raters do not deliberately lie or adjust specific reviews. However, there are many instances in which reviewers have been reported to be paid or bribed to provide a certain rating, from rating games to providing online reviews of 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 an statistically rare number of reviews that deviate from their mean. The second case is in which a given reviewer is bribed on 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. Essentially, the errors in the reviewers' ratings ends up being bimodal, with an unusually high number of high reviews.

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 some other applications where ratings are confined to lie on some finite grid. For instance there are settings where people can just assign scores of 1, 2, 3, 4, 5, or something similar, occasionally with some added half scores. 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 estimating settings in which scores are forced to finite grids is as follows. Consider our model where an expert's rating must be fit into some finite grid. 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 very nonlinear and so the obvious 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 a challenging and important issue for discussion in further research.

References

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.

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.

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

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

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.

Finkelstein, Amy. 2009. "E-Z Tax: Tax Salience and Tax Rates." *Quarterly Journal of Economics* 124:969–1010.

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.

Ginsburgh, Victor A. and Jan C. van Ours. 2003. "Expert Opinion and Compensation: Evidence from a Musical Competition." *The American Economic Review* 93:289–296.

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.

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).*

Reinstein, David A. and Christopher M. Snyder. 2005. "The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics." *Journal of industrial economics* 53:27–51.

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.

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

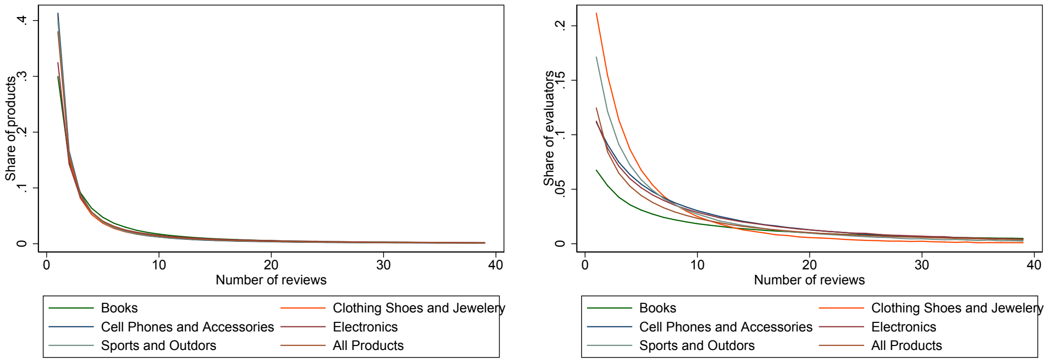
A Amazon ratings data

Table 5: Amazon ratings per product category.

	Total number of			Products					Evaluators	
Product type				Share of products with more than			Median	Mean	with more than one rating	
	Ratings	Evaluators	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 5: 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.

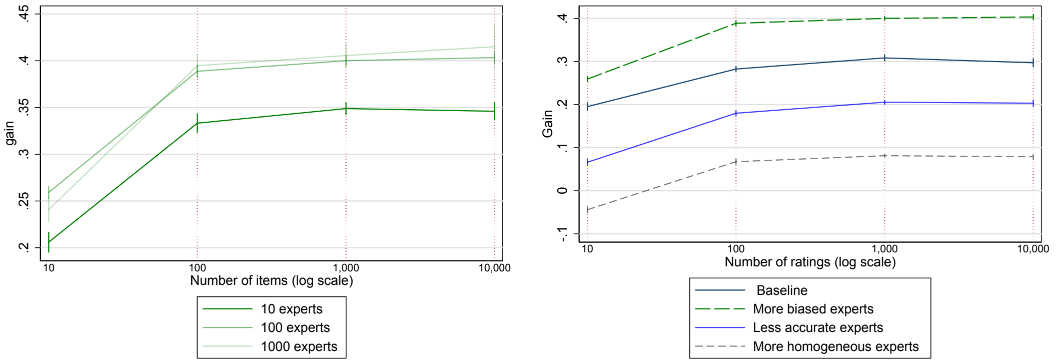


B Proof of Lemma 1

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 (11) 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. ▮

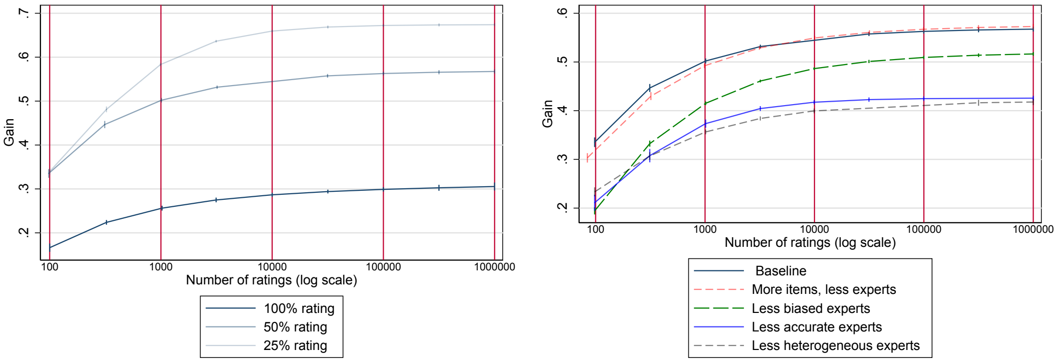
B.1 More Monte Carlo Simulations

Figure 6: Monte Carlo numerical experiments when $f = 1$ (complete rating).



Notes: Fractional polynomial estimates and 95% confidence intervals where each graph point corresponds to 1,000 Monte Carlo data simulations. In the left graph, $f = 1$, $\underline{q} = 0, \overline{q} = 100, \sigma_b = 10, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. In the right graph, the baseline corresponds to $f = 1$, $\underline{q} = 0, \overline{q} = 100, \sigma_b = 20, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. The other series differ from the baseline in the dimensions specified only. The "More biased experts" assumes $\sigma_b = 20$. In the "Less accurate experts" case, we assume $\underline{\sigma} = 10$ and $\overline{\sigma} = 30$, whereas in the "More homogeneous experts" case, $\underline{\sigma} = 10$, and $\overline{\sigma} = 20$.

Figure 7: Monte Carlo numerical experiments, by number of available ratings.



Notes: Fractional polynomial estimates and 95% confidence intervals where each graph point corresponds to 1,000 Monte Carlo data simulations. For each data point, the number of experts and the number of items are adjusted to match the corresponding number of observations according to the horizontal axis. In the left graph, $n = m$, $\underline{q} = 0, \overline{q} = 100, \sigma_b = 10, \underline{\sigma} = 5$, and $\overline{\sigma} = 25$. The percentages indicated in the legends correspond to the proportion f of cells in the rating matrix that are documented. In the right graph, the baseline uses exactly the same parameters as in the left graph with $f = .5$. The other series differ from the baseline in the specified dimensions only. The "More items less experts" case assumes that $n = 10m$. The "Less biased experts" assumes $\sigma_b = 5$. In the "Less accurate experts" case, we assume $\underline{\sigma} = 10$ and $\overline{\sigma} = 30$, whereas in the "Less heterogeneous experts" case, $\underline{\sigma} = 10$, and $\overline{\sigma} = 20$.

C Additional Analysis of Bordeaux Wines and Experts' Ratings

C.1 More on the Data and Estimations

Table 6: 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 7: 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 8: Time distribution of ratings and wine/vintages.

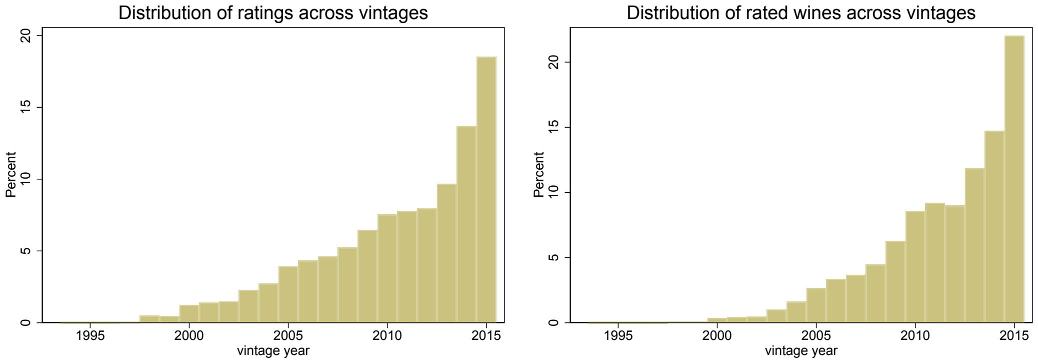


Figure 9: The distribution of the ratings per expert. Though all ratings have been renormalized over a 100-points scale, left graph experts have raw ratings on a 100-points scale initially while right graph experts have a raw rating scale on a 20-points scale.

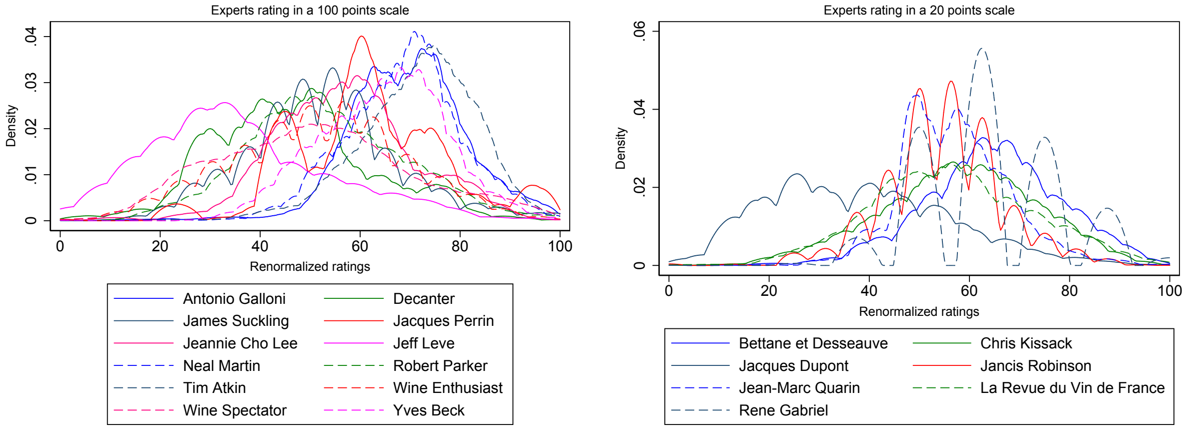


Figure 10: Examples of a two experts' rescaled ratings.

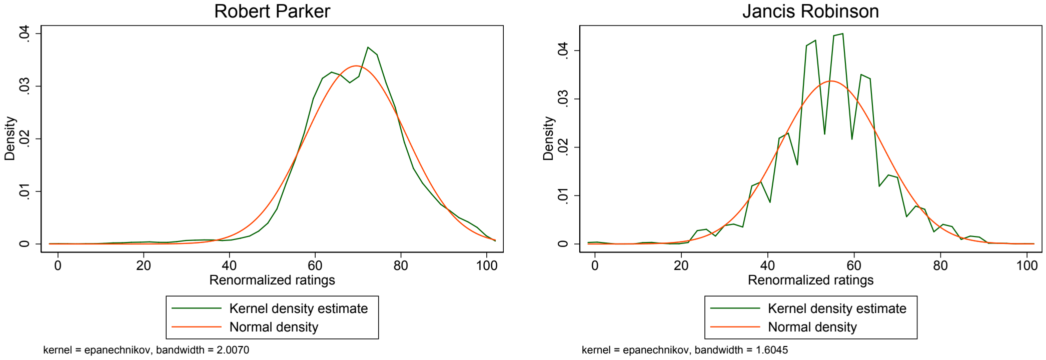


Figure 11: The biases of the experts.

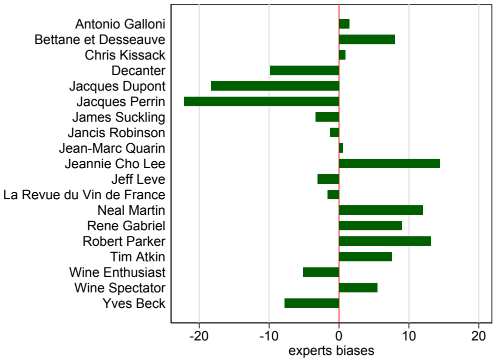


Figure 12: The accuracies of the experts (left graph) and the correlation of their ratings with the estimated wine qualities (right graph).

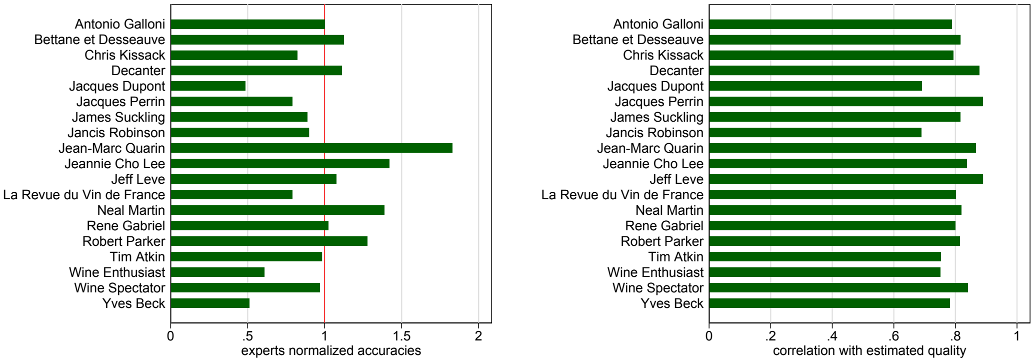


Figure 13: The relationship between the accuracies of experts and the correlation of their ratings with the estimated wine qualities.

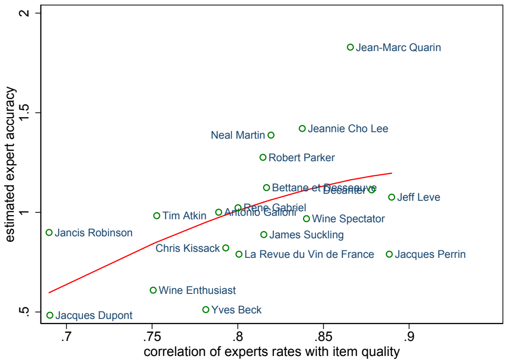


Figure 14: The rescaling of our estimated qualities to adopt the “Parker scale”.

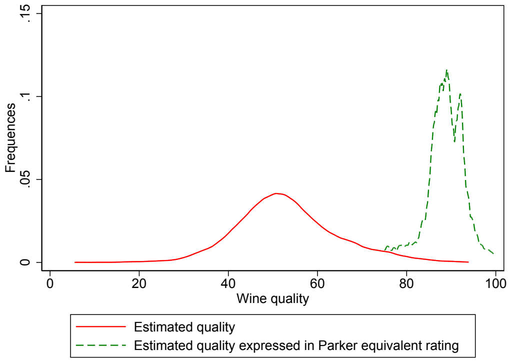


Table 8: The top-100 rated Bordeaux wines.

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

C.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:

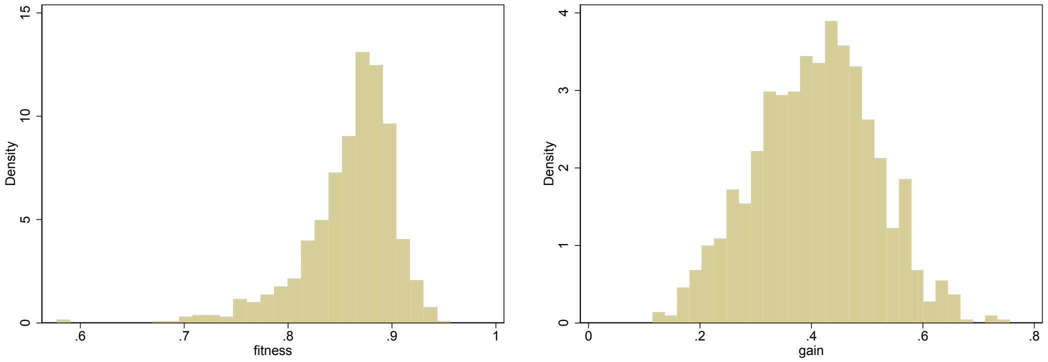
$$\text{Fitness} = 1 - \frac{E\left[(q_i^{two} - q_i)^2\right]}{E\left[(g_{ij} - q_i)^2\right]} = 1 - \frac{E\left[(q_i^{two} - q_i)^2\right]}{E\left[(b_i + \varepsilon_{ij})^2\right]}. \tag{23}$$

Table 9: Descriptive statistics of Fitness and Gain calculated on 1,000 Monte Carlo numerical experiments calibrated on Bordeaux Wine Data

Stats	Fitness	Gain
mean	.86	.41
median	.87	.41
sd	.04	.11
min	.58	.11
max	.96	.76

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 $\overline{\sigma} = 12.914$.

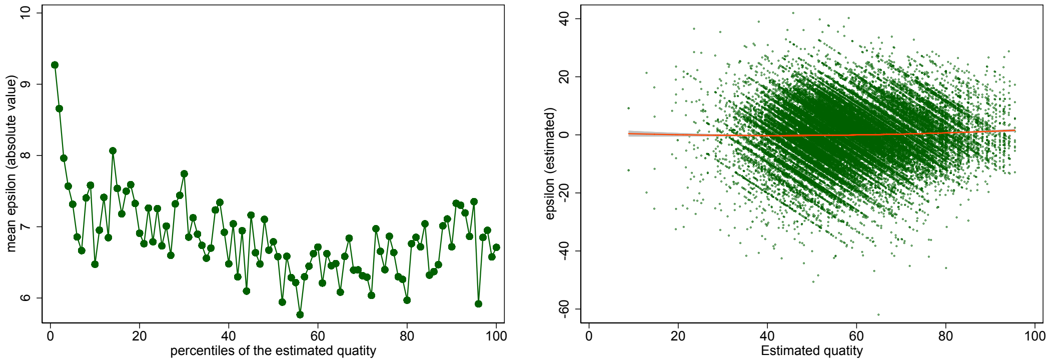
Figure 15: Histograms of Fitness and Gain statistics calculated on 1,000 Monte Carlo numerical experiments calibrated on Bordeaux Wine Data



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 $\overline{\sigma} = 12.914$.

C.3 Errors and Quality for Bordeaux Wines

Figure 16: The relation between percentiles of (estimated) quality (corrected from the expert bias) and experts' (estimated) errors.



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

Table 10: Experts accuracies and biases ranking of Bordeaux red wines only.

Expert	$\left(\sigma_j^{two}\right)^2$	A^{two}	Corr $\left(g_{ij}, q_i^{two}\right)$	b_j^{two}	n_j
Antonio Galloni	73.71	0.99	0.78	1.82	991
Bettane et Desseauve	61.78	1.18	0.82	8.50	2,574
Chris Kissack	89.74	0.82	0.79	0.47	1,983
Decanter	61.32	1.19	0.88	-9.99	1,961
Jacques Dupont	141.80	0.52	0.72	-17.11	2,600
Jacques Perrin	106.99	0.68	0.88	-19.86	427
James Suckling	77.54	0.94	0.82	-3.23	1,722
Jancis Robinson	87.44	0.84	0.68	2.58	3,100
Jean-Marc Quarin	43.62	1.68	0.89	-9.04	2,473
Jeannie Cho Lee	54.29	1.35	0.83	14.95	1,050
Jeff Leve	66.64	1.10	0.88	-3.06	1,408
La Revue du Vin de France	83.21	0.88	0.81	-1.12	1,814
Neal Martin	53.30	1.37	0.82	12.26	2,457
Rene Gabriel	67.85	1.08	0.80	9.04	4,058
Robert Parker	58.63	1.25	0.81	13.23	2,547
Tim Atkin	77.18	0.95	0.75	7.79	1,583
Wine Enthusiast	112.32	0.65	0.77	-5.00	2,050
Wine Spectator	74.70	0.98	0.84	5.28	3,087
Yves Beck	134.47	0.54	0.78	-7.53	394

Figure 17: The accuracies of the experts (left graph) and the correlation of their ratings with the estimated red wine qualities (right graph).

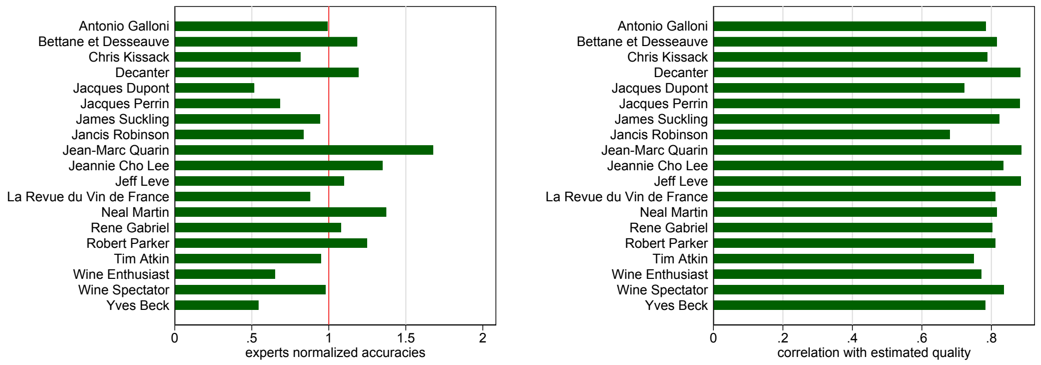
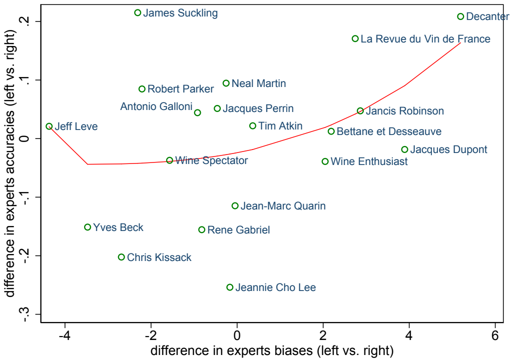


Table 11: The top-100 rated Bordeaux red wines.

Rank	$\widehat{q_j}$	Rescaled	Wine	Vintage	Appellation	Classement
1	93.65	99.5	Margaux	2010	Margaux	Premier Cru Classe en 1855
2	92.62	99.5	Margaux	2015	Margaux	Premier Cru Classe en 1855
3	92.60	99.5	Margaux	2005	Margaux	Premier Cru Classe en 1855
4	92.36	99.5	Grand Vin de Latour	2009	Pauillac	Premier Cru Classe en 1855
5	92.35	99.5	Margaux	2009	Margaux	Premier Cru Classe en 1855
6	91.89	99.5	Petrus	2015	Pomerol	Grands Pomerol
7	91.46	99.5	Grand Vin de Latour	2010	Pauillac	Premier Cru Classe en 1855
8	90.82	99.5	Ausone	2015	Saint Emilion Grand Cru	Premier Cru Classe A
9	90.20	99.5	Grand Vin de Latour	2003	Pauillac	Premier Cru Classe en 1855
10	90.15	99.5	Ausone	2005	Saint Emilion Grand Cru	Premier Cru Classe A
11	90.06	99.5	Lafite Rothschild	2010	Pauillac	Premier Cru Classe en 1855
12	89.95	99.5	Lafite Rothschild	2009	Pauillac	Premier Cru Classe en 1855
13	89.74	99.5	Haut Brion	2009	Pessac Leognan	Premier Cru Classe en 1855
14	89.17	99.5	Cheval Blanc	2015	Saint Emilion Grand Cru	Premier Cru Classe A
15	89.11	99	Haut Brion	2015	Pessac Leognan	Premier Cru Classe en 1855
16	89.06	99	Petrus	2009	Pomerol	Grands Pomerol
17	88.56	99	Ausone	2009	Saint Emilion Grand Cru	Premier Cru Classe A
18	88.50	99	La Mission Haut Brion	2000	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
19	88.47	99	Grand Vin de Latour	2005	Pauillac	Premier Cru Classe en 1855
20	88.44	99	Lafleur	2015	Pomerol	Grands Pomerol
21	88.39	99	Cheval Blanc	2010	Saint Emilion Grand Cru	Premier Cru Classe A
22	88.27	99	Petrus	2010	Pomerol	Grands Pomerol
23	88.27	99	Lafite Rothschild	2005	Pauillac	Premier Cru Classe en 1855
24	87.48	99	Lafite Rothschild	2003	Pauillac	Premier Cru Classe en 1855
25	87.29	99	Cheval Blanc	2009	Saint Emilion Grand Cru	Premier Cru Classe A
26	87.26	99	Grand Vin de Latour	2000	Pauillac	Premier Cru Classe en 1855
27	87.13	99	Haut Brion	2010	Pessac Leognan	Premier Cru Classe en 1855
28	86.84	99	Haut Brion	2005	Pessac Leognan	Premier Cru Classe en 1855
29	86.51	99	Lafleur	2009	Pomerol	Grands Pomerol
30	86.43	99	Petrus	2005	Pomerol	Grands Pomerol
31	86.25	99	Leoville Las Cases	2009	Saint Julien	Deuxieme Cru Classe en 1855
32	86.10	99	Mouton Rothschild	2009	Pauillac	Premier Cru Classe en 1855
33	86.10	99	Eglise Clinet	2009	Pomerol	Grands Pomerol
34	86.05	99	Vieux Chateau Certan	2010	Pomerol	Grands Pomerol
35	85.88	99	Cheval Blanc	2005	Saint Emilion Grand Cru	Premier Cru Classe A
36	85.85	99	Grand Vin de Latour	2015	Pauillac	Premier Cru Classe en 1855
37	85.78	99	Mouton Rothschild	2010	Pauillac	Premier Cru Classe en 1855
38	85.15	99	Ausone	2010	Saint Emilion Grand Cru	Premier Cru Classe A
39	85.13	99	Montrose	2003	Saint Estephe	Deuxieme Cru Classe en 1855
40	85.09	99	Canon	2015	Saint Emilion Grand Cru	Premier Cru Classe B
41	84.85	99	La Mission Haut Brion	2015	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
42	84.72	99	Cos d'Estournel	2003	Saint Estephe	Deuxieme Cru Classe en 1855
43	84.64	99	Mouton Rothschild	2015	Pauillac	Premier Cru Classe en 1855
44	84.45	99	Palmer	2015	Margaux	Troisieme Cru Classe en 1855
45	84.42	99	Ausone	2003	Saint Emilion Grand Cru	Premier Cru Classe A
46	84.39	99	Vieux Chateau Certan	2015	Pomerol	Grands Pomerol
47	84.16	99	Pavie	2000	Saint Emilion Grand Cru	Premier Cru Classe A
48	84.15	98	Palmer	2009	Margaux	Troisieme Cru Classe en 1855
49	84.05	98	Lafleur	2005	Pomerol	Grands Pomerol
50	83.96	98	Lafleur	2010	Pomerol	Grands Pomerol
51	83.82	98	Grand Vin de Latour	2004	Pauillac	Premier Cru Classe en 1855
52	83.79	98	Leoville Las Cases	2005	Saint Julien	Deuxieme Cru Classe en 1855
53	83.63	98	Petrus	1998	Pomerol	Grands Pomerol
54	83.61	98	La Mission Haut Brion	2010	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
55	83.56	98	Angelus	2015	Saint Emilion Grand Cru	Premier Cru Classe A
56	83.55	98	Leoville Las Cases	2000	Saint Julien	Deuxieme Cru Classe en 1855
57	83.47	98	Cheval Blanc	2000	Saint Emilion Grand Cru	Premier Cru Classe A
58	83.41	98	Eglise Clinet	2010	Pomerol	Grands Pomerol
59	83.38	98	Pontet Canet	2009	Pauillac	Cinquieme Cru Classe en 1855
60	83.30	98	Margaux	2003	Margaux	Premier Cru Classe en 1855
61	82.97	98	Trotanoy	2009	Pomerol	Grands Pomerol
62	82.72	98	Haut Bailly	2015	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
63	82.56	98	Leoville Barton	2000	Saint Julien	Deuxieme Cru Classe en 1855
64	82.51	98	Ausone	2008	Saint Emilion Grand Cru	Premier Cru Classe A
65	82.44	98	Figeac	2015	Saint Emilion Grand Cru	Premier Cru Classe B
66	82.43	98	Troplong Mondot	2005	Saint Emilion Grand Cru	Premier Cru Classe B
67	82.38	98	Petrus	2012	Pomerol	Grands Pomerol
68	82.35	97.5	Eglise Clinet	2015	Pomerol	Grands Pomerol
69	82.22	97.5	Lafite Rothschild	2015	Pauillac	Premier Cru Classe en 1855
70	82.14	97.5	Montrose	2009	Saint Estephe	Deuxieme Cru Classe en 1855
71	82.14	97.5	Vieux Chateau Certan	2009	Pomerol	Grands Pomerol
72	82.12	97.5	Trotanoy	2015	Pomerol	Grands Pomerol
73	82.02	97.5	La Mission Haut Brion	2009	Pessac Leognan	Grand Cru Classe de Graves (Rouge)
74	82.00	97.5	Leoville Las Cases	2015	Saint Julien	Deuxieme Cru Classe en 1855
75	81.98	97.5	Grand Vin de Latour	2014	Pauillac	Premier Cru Classe en 1855
76	81.92	97.5	Cos d'Estournel	2010	Saint Estephe	Deuxieme Cru Classe en 1855
77	81.91	97.5	Leoville Las Cases	2010	Saint Julien	Deuxieme Cru Classe en 1855
78	81.84	97.5	Palmer	2010	Margaux	Troisieme Cru Classe en 1855
79	81.79	97.5	Trotanoy	1998	Pomerol	Grands Pomerol
80	81.74	97.5	Palmer	2005	Margaux	Troisieme Cru Classe en 1855
81	81.59	97.5	Ducru Beaucaillou	2009	Saint Julien	Deuxieme Cru Classe en 1855
82	81.37	97.5	Pavie	2015	Saint Emilion Grand Cru	Premier Cru Classe A
83	81.27	97.5	Mouton Rothschild	2006	Pauillac	Premier Cru Classe en 1855
84	81.14	97.5	Trotanoy	2010	Pomerol	Grands Pomerol
85	81.11	97.5	Ducru Beaucaillou	2010	Saint Julien	Deuxieme Cru Classe en 1855
86	81.08	97.5	Ducru Beaucaillou	2015	Saint Julien	Deuxieme Cru Classe en 1855
87	80.84	97.5	Leoville Las Cases	2006	Saint Julien	Deuxieme Cru Classe en 1855
88	80.68	97.5	Margaux	2006	Margaux	Premier Cru Classe en 1855
89	80.51	97.5	Pontet Canet	2010	Pauillac	Cinquieme Cru Classe en 1855
90	80.51	97.5	Cos d'Estournel	2005	Saint Estephe	Deuxieme Cru Classe en 1855
91	80.45	97.5	Pichon Baron	2010	Pauillac	Deuxieme Cru Classe en 1855
92	80.40	97.5	Mouton Rothschild	2002	Pauillac	Premier Cru Classe en 1855
93	80.40	97.5	Lafite Rothschild	2000	Pauillac	Premier Cru Classe en 1855
94	80.32	97.5	Tertre Roteboeuf	2015	Saint Emilion	
95	80.17	97.5	Le Pin	2010	Pomerol	Grands Pomerol
96	79.98	97.5	Pichon Comtesse	2015	Pauillac	Deuxieme Cru Classe en 1855
97	79.96	97.5	Vieux Chateau Certan	1998	Pomerol	Grands Pomerol
98	79.96	97.5	Haut Brion	1998	Pessac Leognan	Premier Cru Classe en 1855
99	79.94	97.5	Ausone	2014	Saint Emilion Grand Cru	Premier Cru Classe A
100	79.83	97	Evangile	2005	Pomerol	Grands Pomerol

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

Figure 18: The relationship between the differences in accuracies and the differences in biases (between left bank and right bank wines).

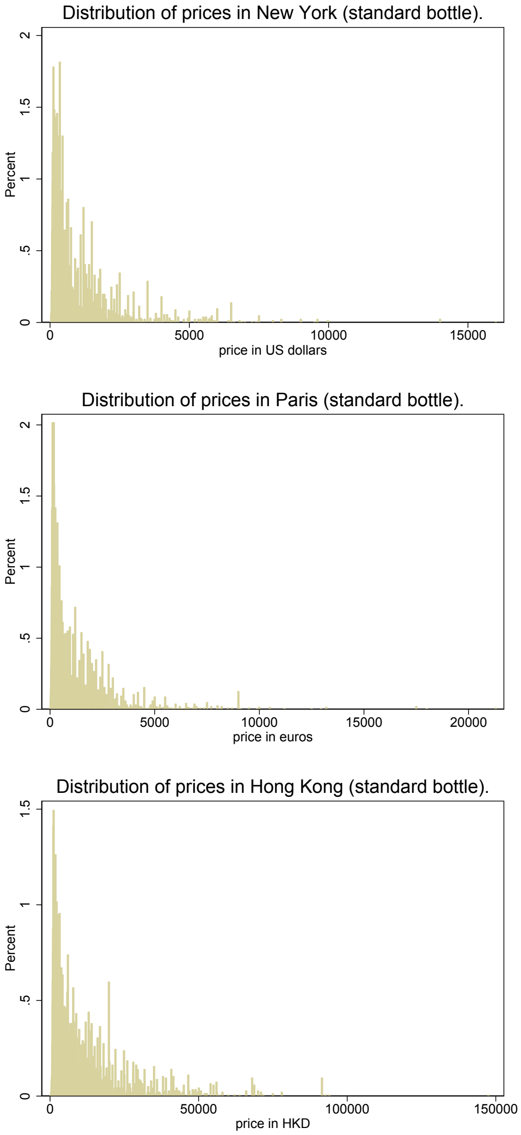


D More on Prices and Ratings of Bordeaux Wines

Table 12: 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 19: Prices in the three markets (in local currency).



Distribution of prices in New York (standard bottle).

Distribution of prices in Paris (standard bottle).

Distribution of prices in Hong Kong (standard bottle).

Table 13: Top-100 most surveyed stores (restaurants).

Store	Market	Number of Wines	Number of Prices
L'Atelier de Joel Robuchon - HK	Hong Kong	429	1,607
La Truffiere	Paris	409	1,270
Le Cinq - Paris	Paris	288	581
Le Carre des Feuillants	Paris	272	1,077
Apicius	Paris	272	397
Le Pre Catelan	Paris	263	431
Petrus - HK	Hong Kong	237	917
Epicure	Paris	234	370
Cepage	Hong Kong	223	507
L Abeille (Shangri-La)	Paris	190	558
Per Se	New York	172	265
KO Dining Group (Messina, Yu Lei, Kazuo Okuda)	Hong Kong	171	608
Mandarin Oriental Paris - Sur Mesure, Camelia	Paris	159	411
Le Meurice	Paris	156	410
21 Club	New York	154	394
Shang Palace (Shangri-La) - Paris	Paris	154	282
Au Trou Gascon	Paris	147	505
The Steak House winebar + grill	Hong Kong	137	321
Alain Ducasse au Plaza Athenee	Paris	136	326
Spoon	Hong Kong	136	281
Le relais du plaza (plaza athenee)	Paris	132	149
Le Grand Vefour	Paris	131	276
Yan Toh Heen	Hong Kong	129	241
The Modern	New York	129	180
Aureole	New York	128	246
Amber	Hong Kong	125	191
Blt Steak	New York	124	171
Le Diane	Paris	118	232
Pierre - HK	Hong Kong	116	288
Fouquet's	Paris	115	180
Sparks Steak House	New York	115	407
Mandarin Bar and Grill	Hong Kong	115	246
Tin Lung Heen	Hong Kong	113	182
Daniel	New York	112	277
Man Wah	Hong Kong	107	221
Eleven Madison Park	New York	107	174
Morrell Wine Bar & Cafe	New York	104	144
City Winery	New York	103	151
Shang Palace - HK	Hong Kong	102	356
Porter House	New York	101	147
Jean Georges	New York	101	136
Veritas	New York	98	255
Asiate	New York	96	191
Jean-Francois Piege	Paris	95	95
Le Cirque	New York	91	129
Mathieu Pacaud - Histoires	Paris	91	91
Pierre Gagnaire	Paris	91	129
Conrad Hotel (Golden Leaf)	Hong Kong	91	152
Hexagone	Paris	90	90
Caprice	Hong Kong	89	284
The Mark Restaurant by Jean-Georges	New York	88	134
Benoit - Paris	Paris	88	114
Cafe Boulud New York	New York	83	131
G Bar	Hong Kong	83	178
Le Gabriel - Paris	Paris	81	81
Harlan's	Hong Kong	81	160
Le Bernardin	New York	81	111
Gordon Ramsay Au Trianon	Paris	81	81
Sevva	Hong Kong	81	81
Pur'	Paris	80	121
Guy Savoy	Paris	79	79
Chez Flottes	Paris	79	153
Tosca - HK	Hong Kong	79	143
L'Altro - HK	Hong Kong	79	168
Bouley	New York	78	105
Picholine	New York	77	99
A Voce - Columbus	New York	77	181
Hotel Park Hyatt- Paris Vendome	Paris	76	101
Angelini	Hong Kong	76	76
Nice Matin	New York	76	181
Lili au Peninsula	Paris	73	115
Ming Court	Hong Kong	73	218
La Compagine des Vins surnaturels	Paris	73	83
Fook Lam Moon - Hong Kong	Hong Kong	73	141
Drouant	Paris	72	90
Bibo	Hong Kong	71	71
Blt Prime - NYC	New York	71	121
La Table du Lancaster	Paris	70	127
Le Violon d'Ingres	Paris	70	91
NOBU Intercontinental Hong Kong	Hong Kong	70	171
Gabriel Kreuther	New York	69	69
Michel Rostang	Paris	68	68
L'Atelier de Joel Robuchon - Paris	Paris	68	99
Cuisine Cuisine at Mira	Hong Kong	68	135
Smith & Wollensky New York	New York	68	96
Mandarin Oriental (Krug Room)	Hong Kong	68	101
Cuisine Cuisine at IFC	Hong Kong	68	75
Le Beef Club / Fish Club	Paris	67	84
La Grande Cascade	Paris	67	67
Nicholini's	Hong Kong	67	70
Dominique Bouchet	Paris	66	93
Gotham Bar and Grill	New York	64	100
Benoit - New York	New York	64	86
Lung King Heen	Hong Kong	64	163
Les 110 de Taillevent	Paris	63	118
Rouge Tomate	New York	63	120
Harrys Cafe and Steak	New York	63	128
Le Celadon	Paris	63	108
La Scene - Hotel Prince de Galles	Paris	62	62

Table 14: Retail prices as a function of each expert "en primeur" ratings.

	coef	t-stat	N	r2	aic	bic
Bettane et Desseauve	1.707^{+}	(12.12)	21,664	0.756	29388.7	30825.8
Chris Kissack	0.884^{+}	(5.77)	10,059	0.730	12120.5	13267.8
Decanter	1.155^{+}	(9.48)	4,241	0.794	4362.7	5023.3
Jacques Dupont	0.541^{+}	(8.64)	23,898	0.753	33457.4	35106.1
Jancis Robinson	1.198^{+}	(13.43)	28,774	0.766	40215.6	42232.8
Jean-Marc Quarin	2.536^{+}	(13.97)	21,645	0.779	27484.4	29041.0
Neal Martin	1.119^{+}	(5.54)	15,523	0.772	19186.7	20548.4
Rene Gabriel	1.396^{+}	(13.83)	37,285	0.756	55249.5	57585.7
La Revue du Vin de France	1.057^{+}	(9.02)	14,118	0.771	17906.8	19115.7
Robert Parker	1.956^{+}	(7.54)	36,109	0.755	53433.0	55624.5
Tim Atkin	1.071^{+}	(6.31)	3,304	0.811	3627.5	4250.0
Wine Enthusiast	0.994^{+}	(11.56)	16,636	0.792	22813.1	24125.4
Wine Spectator	1.170^{+}	(14.01)	38,917	0.751	58696.9	61027.7

Notes: Column "coef" exhibits estimated coefficients of each expert ratings in a linear regression on wines retail prices. *t*-statistics are in parentheses. 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 16). Prices and ratings are in log so that coefficients can be interpreted as elasticities. Regressions converged only for experts who have rated wines for which we have a sufficient number of prices (more than 3000) given the high number of fixed effects considered. It did not converge for Antonio Galloni, Jacques Perrin, James Suckling, Jeannie Cho Lee, Jeff Leve, Yves Beck.

Table 15: 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, re-rating year, vintage×appellation, type (color), and official ranking. Ratings are corrected to span the 0-100 scale (see Equation 16). Prices and ratings are in log so that coefficients can be interpreted as elasticities. Only the experts who have rated wines for which we have a sufficient number of prices (>2000) are considered here.

Table 16: 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*	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, re-rating year, vintage×appellation and official ranking. Ratings are corrected to span the 0-100 scale (see Equation 16). Prices and ratings are in log so that coefficients can be interpreted as elasticities. Only the experts who have rated wines for which we have a sufficient number of prices (>2000) are considered here.

Table 17: 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, re-rating year, vintage×appellation and official ranking. Ratings are corrected to span the $0-100$ scale (see Equation 16). Prices and ratings are in log so that coefficients can be interpreted as elasticities. Only the experts who have rated wines for which we have a sufficient number of prices ($> 2{,}000$) are considered here.

Table 18: Retail prices as a function of "en primeur" ratings by the top-5 most influential experts. 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, re-rating year, vintage×appellation and official ranking. Ratings are corrected to span the 0-100 scale (see Equation 16). Prices and ratings are in log so that coefficients can be interpreted as elasticities. Only the experts who have rated wines for which we have a sufficient number of prices (>2000) are considered here.

E Estimated Qualities and the Re-Rating of Bordeaux Wines

E.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 19 in Online Appendix E.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.

Online Appendix Section F.2 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 19: Summary statistics on the re-rating data, by expert.

Expert	# Re-ratings	# Re-rated wines	Av. Initial rating	Av. Adjustment
Decanter	6	5	55.03	16.88
James Suckling	585	499	61.82	11.14
Jancis Robinson	5,173	2,074	60.11	10.16
Neal Martin	1,380	824	63.25	10.94
Robert Parker	3,307	2,105	62.23	10.75
Wine Spectator	2,288	2,157	61.76	18.86

E.2 Results

The results are presented in Table 20. In the first column, the rerating is regressed only on estimated quality, on the top of all fixed effects. In the 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.

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

In Table 21, 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 regress remain. These regressions show that experts adjust their ratings about 21 to 30 percent

Table 20: 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 16). All ratings are in log.

(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 21: 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 16).

F 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.

F.1 Prices

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

$$q = f + \phi. \tag{24}$$

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

$$g^r = E(q|\, s^r, f) = f + E(\phi|\, s) = f + s^r, \tag{25}$$

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

$$\begin{aligned} E(q|g,f,S) &= \lambda E(q|f,S)+\mu a+(1-\lambda-\mu)\left(E(q|s^r,f)\right) \\ &= \lambda\widehat{q}+\mu a+(1-\lambda-\mu)g^r \end{aligned} \tag{26}$$

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:

$$p = \beta\widehat{q}+\beta^r g^r+\nu_a+\nu_f+\nu_t+\nu_{sto}+\varepsilon, \tag{27}$$

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 (27) 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.

F.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 25 (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:

$$g' = E(q|s,s',s^r,f) = f+E(\phi|s,s',s^r). \tag{28}$$

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 (28) becomes

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

Using Equations 24 and 25, 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:

$$g' = \beta_1 \widehat{q} + \beta_2 g + \beta_3 g^r + \nu_a + \nu_f + \nu_t + \nu_e + \epsilon', \tag{29}$$

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.