Standardizing Expert Wine Scores: An Application for Bordeaux en primeur

Jean-Marie Cardebat, Emmanuel Paroissien

Standardizing Expert Wine Scores: An Application for Bordeaux *en primeur**

Jean-Marie Cardebat [a] **and Emmanuel Paroissien** [b]

Abstract

In this paper we provide a simple and transparent non parametric methodology to express the scores of each wine expert on the same rating scale. We discuss the advantage of this methodology over a linear transformation. The non-paramatric method ensures the comparability of scores among experts and allows for a relevant average calculation of available wine scores. (JEL Classifications: Q13, L15, C14).

Keywords: Wine, quality, experts.

I. Introduction

As an experience good, the quality of a wine is only known after its consumption. In contrast to consumers, wine producers are informed about their products' quality. This information asymmetry has led to the emergence of wine experts providing information on wine quality. The contingent information market is particularly well-developed in the wine sectors where numerous experts coexist. The subjectivity of the wine quality assessment, the regional segmentations,[1] or their (supposed) preferences (Storchmann, 2012) partly justify a large number of experts. Moreover, the grading systems and habits could differ from one expert to another. In particular, the European experts are used to rating wine on a 20-point scale whilst US experts

* We thank an anonymous referee for his/her comments that helped to improve this paper. All remaining errors are ours.
[a] Larefi, University of Bordeaux, Avenue Léon Duguit, 33608 Pessac, France, and INSSEC Bordeaux, France; e-mail: jean-marie.cardebat@u-bordeaux.fr
[b] Larefi, University of Bordeaux, Avenue Léon Duguit, 33608 Pessac, France; e-mail: emmanuel.paroissien@u-bordeaux.fr
[1] By regional segmentation we refer to the fact that not only are experts more or less specialized in wines coming from specific regions, but also that some experts target specific consumers (at least as regards the choice of the language in which they edit their comments).

use 100 points (e.g., Masset and Weisskopf, 2015). The heterogeneity of the rating systems can increase the consumer's perceived uncertainty. The question of rating homogenization on the same scale of preferences is therefore at the heart of the uncertainty debate about wine quality.

The uncertainty about wine quality is particularly high during the *en primeur* campaign in the Bordeaux Region. The *primeur* market can be seen as a forward market dedicated to fine Bordeaux wines. The *en primeur* campaign takes place during the spring, starting with a huge multi-day tasting organized by the chateaux in the first week of April. Wine merchants, wine enthusiasts, and of course, wine experts are involved in this event. They all taste the wine from the latest harvest. Therefore, the wine is not yet vinified and the quality assessment is particularly difficult and uncertain. The aim of this campaign is to sell (chateaux) and buy (wine merchants)[2] before the wine is effectively released in bottles, which will happen about 18 months later. The prices and quantities exchanged are determined during the *en primeur* campaign and the wine will be delivered once it is bottled.

The economic stakes of the tasting are therefore extremely high because prices and quantities exchanged are influenced by the experts' scores. The wine economics literature has provided ample evidence of the link between *en primeur* wine prices and the experts' scores (see notably Hadj Ali and Nauges, 2007; Hadj Ali et al., 2008; Masset et al., 2015). Another strand of the literature deals with the information contained in the experts' grades (see for example Ashenfelter et al., 1995; Ashenfelter, 2008; or, more recently, Cardebat et al., 2014), the divergence between experts (notably Ashton, 2012, 2013; Hodgson, 2008; Masset et al., 2015; Olkin et al., 2015) or the randomness of the tastings (e.g., Ashton, 2014; Quandt, 2007; Bodington, 2015).

However, no paper has tried to express the experts' scores on the same scale of preference or in the same rating system before analyzing the grades divergence or bias or impact on prices. As noted by Masset et al. (2015, p.80) "Comparisons are difficult to make, as not all experts use the same scale to establish their scores". Furthermore, as far as we know, there is no paper trying to provide a global score aggregating all the marks released by experts during the *en primeur* campaign, although a demand exists for such a global score from the professionals. However, if no academic papers exist, in the wine industry, most of the web merchants provide such aggregated scores (see, for example, wine decider or wine searcher). The website of Bertrand Leguern is also dedicated to the calculation of an aggregated score which is used by wine professionals. Nevertheless, we cannot find any information on the way these scores have been calculated. There is no transparency in their calculation, thereby reinforcing the information asymmetry instead of reducing it.

[2] The wine merchants (called *negociants* in Bordeaux) are free to buy or not, but they receive allocations (the right to buy in a certain amount) from the chateaux and if they do not buy a specific year, the chateaux may remove their allocations for the following year.

Wine professionals, particularly the *negociants* who buy *en primeur* wines, request aggregated and transparent information on wine quality rather than comparing numerous grades emanating from avariety of experts. This highlights the importance of reducing the information asymmetry and therefore increasing the *en primeur* market efficiency (Mahenc and Meunier, 2006). Given the pending retirement of the main expert, Robert Parker, harmonizing experts' scores appears particularly useful since Parker's disappearance will reinforce the uncertainty and the need for a reference score.

The aim of this paper is, therefore, to develop a methodology for calculating a single score aggregating the grades released by 15 experts who have traditionally been scoring Bordeaux *en primeur* wines since the beginning of the last decade. Based on a large database of Bordeaux *en primeur* expert scores, we suggest a methodology to translate the rating scale of one expert into the rating scale of another, thereby facilitating the comparability of all the experts' scores. The global score is then basically calculated as a simple arithmetic average of these transformed scores. This aggregated score has the potential to be considered as a new reference score on the fine wine market.

This study may be interesting to academics who may benefit from a methodology ensuring proper expert score comparisons by taking into account the different rating systems among experts. In addition, based on this methodology, we provide wine professionals with a unique standardized wine score aggregating the information coming from all experts operating on the *en primeur* market.

The remainder of this paper is structured as follows: the next sections present our dataset, while section III. displays the methodology of the standardized wine score; section IV. reports the standardized scores and discusses the results following different robustness checks; the last section concludes.

II. Data

Our dataset contains the scores given by 15 well-known wine experts[3] during the *en primeur* campaign over the period from 2000 to 2014. All the wines rated by these experts are present in the dataset which represents 447 chateaux and 4333 chateau-vintage pairs; that is, on average, each chateau is rated 9.7 times over the observed period.

[3] The term "expert" is used here indifferently to designate a person (James Suckling, Jancis Robinson, etc.) or an organization (i.e., magazines like Wine Spectator or La Revue du Vin de France – RVF, etc.). Decanter has a special status in the sense that we split its scores into two categories: *Decanter 20* and *Decanter 100* because Decanter chose to change its traditional 20-point scale for a 100-point scale during the period studied. We have therefore decided to consider its scores on a 20-point scale and on a 100-point scale as two different experts.

Table 1
Descriptive Statistics of Expert Score Data

Expert	*Frequency*	*Min*	*Max*	*Mean*	*Median*	*Standard Dev.*
Rene Gabriel	3,639	12	20	17.12	17	1.14
Wine Spectator	2,886	77	98.5	90.2	90	3.45
Robert Parker	2,609	71.5	99.5	90.38	90.5	3.52
Jancis Robinson	2,538	12	20	16.4	16.5	0.99
Jacques Dupont	2,156	13	20	15.82	16	1.28
Bettane & Desseauve	2,113	10	20	16.56	16.5	1.33
Neal Martin	1,711	70	99	90.03	90	3.53
Decanter20	1,615	14.5	20	16.93	17	1.04
Jean-Marc Quarin	1,497	10	20	15.74	15.75	1.07
James Suckling	1,059	84.5	100	91.25	91.5	2.72
Decanter100	1,026	81	95	88.19	88	2.91
Tim Atkin	1,011	82	100	91.37	92	3.35
La RVF	484	11.5	20	16.29	16.25	1.38
Jeannie Cho Lee	219	80	99	91.87	92	2.81
Antonio Galloni	210	79	95.5	89.24	89.5	2.77
Jeff Leve	158	83	99	90.33	90	3.01

Source: Authors' calculation based on Wine Services (2015) data.

The first column in Table 1 shows the number of wines effectively rated by each expert. Rene Gabriel appears to be the most productive expert with 3,639 scores over the period. Similarly, five additional experts are highly active on the wine opinion market. They all have rated more than 2000 *en primeur* wines between 2000 and 2014. In contrast, the last four experts of this list exhibit a significantly weaker activity with fewer than 500 scores each. The following columns display the traditional descriptive statistics on the experts' scores. Among the 16 (15 + 1, see footnote 1) experts, seven use a 20-point grading scale, they are all European, and nine use a 100-point scale, they are overwhelmingly American. The Chinese J. Cho Lee and the British Tim Atkin are exceptions.

The scores given by the experts seem relatively homogenous and average between 15.74 and 17.12 for the European raters and between 89.24 and 91.87 for the U.S. experts. Interestingly, we note that the Europeans have all awared a 20-point maximum grade at least once while only J. Suckling and Tim Atkin have handed out the maximum 100-point grade. The score range defined as the difference between the maximum and the minimum score for each expert lies between 14 and 29 for the US experts and 5.5 to 10 for the European experts.

We note two remarkable facts. First, all experts utilize only a fraction of their scale. In comparison, the fraction utilized by U.S. experts seems to be particularly small (20 points on average). However, in absolute values this exceeds the spectrum used by European experts (7.8 points on average), giving the former a potentially higher accuracy in their rating. Second, both U.S. and European raters exhibit significant differences in the way they rate the wines: there is no homogeneity among them concerning score range they use. Therefore, the direct comparison among experts' scores is fallacious, even if they use the same rating scale. Each expert has his/her own preference space and our aim is to express all scores in the same space of preferences.

The medians also offer interesting information as they can be interpreted as a threshold between good wines and less good/bad wines. 90 points (16.5) for the U.S. (European) experts appears to be the dividing line between these two catagories.

Table 2 presents the number of wines that have been tasted by each expert pair, i.e., by at least two experts. With 2698 wines rated both by Ren Gabriel and Wine Spectator, these two experts exhibit the highest overlap. On average, Robert Parker, Neal Martin, Jancis Robinson, Wine Spectator, Bettane & Desseauve, Jacques Dupont, La Revue du Vin de France, and Rene Gabriel have rated more than 1000 identical wines over the observed time period.

Table 3 reports a systematic positive correlation between each expert pair; however, the average correlation among experts does not exceed 0.59. Jean-Marc Quarin and Jeffe Leve exhbit the highest correlation. Jancis Robinson and Antonio Galloni exhibit the lowest correlation and therefore the lowest agreement (concordance) with the other experts. In contrast, Jeff Leve and Decanter 20 display the highest correlation and therefore the best level of concordance with the other experts. In particular, these two experts' grades are strongly correlated with those by Robert Parker. The U.S. experts seem to have higher concordance among themselves compared to the European ones. These results are in line with the work of Masset and Weisskopf (2015), even if their results suggest a high level of concordance among various wine raters. In contrast, given an average correlation of 0.59 and a high volatility of the correlation coefficients, we do not deem the level of expert concordance particularly high.

III. Methodology

Robert Parker and Jancis Robinson are influential experts, in the U.S. and in England, respectively, and best embody the issue of transforming the grading scales. While Robert Parker scores out of 100 points, Jancis Robinson scores out of 20 points. Our method addresses a common quality assessment problem. Imagine a comparison between two wines where the first is graded by both experts, but the second one is only rated by Robert Parker. The key issue how to properly utilize the information given by Jancis Robinson and translate them into Parker scores.

The naïve solution is the linear function by simply multiplying Jancis Robinson's scores by a factor of five. However, this solution is unsatisfactory, as it disregards the utilized score range of [12,20] for Robinson and [70,100] for Parker. In order to consider the minima of the intervals utilized by each expert, one can employ an affine

Table 2

Wine Pairings: Number of Identical Wines Tasted by Two Experts

	RP	*NM*	*JR*	*WS*	*AG*	*BD*	*JD*	*JS*	*JC*	*JL*	*RVF*	*JMQ*	*RG*	*TA*	*D20*	*D100*
RP		1361	1833	2168	168	1637	1667	706	231	568	1317	1041	2443	750	811	198
NM	1361		1549	1419	160	1422	1353	714	243	556	1294	996	1578	787	842	170
JR	1833	1549		2049	171	1929	1946	730	247	561	1440	1268	2361	838	898	194
WS	2168	1419	2049		168	1753	1803	663	221	529	1330	1151	2698	743	832	178
AG	168	160	171	168		161	164	177	1	173	157	167	184	158	13	183
BD	1637	1422	1929	1753	161		1756	687	232	570	1427	1185	2007	773	856	200
JD	1667	1353	1946	1803	164	1756		650	230	515	1316	1139	2039	738	796	186
JS	706	714	730	663	177	687	650		219	535	618	652	867	731	565	211
JC	231	243	247	221	1	232	230	219		158	211	186	250	227	246	1
JL	568	556	561	529	173	570	515	535	158		484	427	599	524	403	203
RVF	1317	1294	1440	1330	157	1427	1316	618	211	484		959	1546	690	756	191
JMQ	1041	996	1268	1151	167	1185	1139	652	186	427	959		1366	657	538	206
RG	2443	1578	2361	2698	184	2007	2039	867	250	599	1546	1366		918	929	214
TA	750	787	838	743	158	773	738	731	227	524	690	657	918		670	181
D20	811	842	898	832	13	856	796	565	246	403	756	538	929	670		0
D100	198	170	194	178	183	200	186	211	1	203	191	206	214	181	0	
Average	1262	1054	1340	1329	154	1231	1215	602	202	473	1008	878	1495	656	653	180

Source: Authors' calculation based on Wine Services (2015) data. WS: Wine Spectator; RP: Robert Parker; JR: Jancis Robinson; JD: Jacques Dupont; BD: Bettane & Desseauve; NM: Neal Martin; D20: Decanter20; JS: James Suckling; D100: Decanter100; RVF: La Revue du Vin de France; JCL: Jeannie Cho Lee; AG: Antonio Galloni; JL: Jeff Leve; JMQ: Jean-Marc Quarin; TA: Tim Atkin; RG: Rene Gabriel.

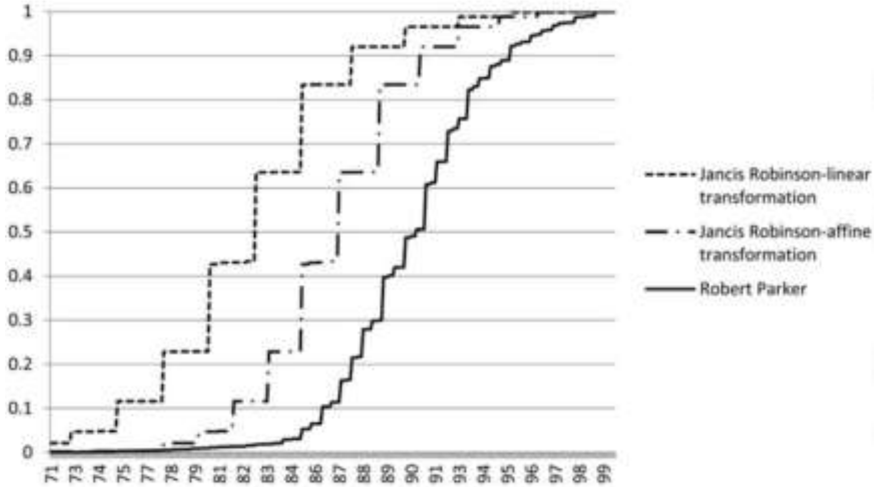
Table 3
Expert Score Correlation Matrix

	RP	*NM*	*JR*	*WS*	*AG*	*BD*	*JD*	*JS*	*JC*	*JL*	*RVF*	*JMQ*	*RG*	*TA*	*D20*	*D100*
RP		0.57	0.43	0.61	0.41	0.59	0.50	0.69	0.68	0.77	0.60	0.68	0.58	0.57	0.71	0.58
NM	0.57		0.49	0.62	0.56	0.58	0.50	0.69	0.59	0.74	0.59	0.65	0.58	0.57	0.67	0.61
JR	0.43	0.49		0.51	0.17	0.50	0.39	0.48	0.55	0.42	0.46	0.54	0.45	0.57	0.63	0.36
WS	0.61	0.62	0.51		0.59	0.62	0.47	0.74	0.70	0.75	0.60	0.64	0.61	0.64	0.69	0.62
AG	0.41	0.56	0.17	0.59		0.45	0.35	0.47		0.60	0.44	0.52	0.54	0.35	0.32	0.56
BD	0.59	0.58	0.50	0.62	0.45		0.50	0.65	0.70	0.67	0.65	0.69	0.55	0.63	0.75	0.74
JD	0.50	0.50	0.39	0.47	0.35	0.50		0.59	0.60	0.62	0.56	0.62	0.47	0.52	0.63	0.67
JS	0.69	0.69	0.48	0.74	0.47	0.65	0.59		0.66	0.75	0.70	0.70	0.66	0.54	0.71	0.60
JC	0.68	0.59	0.55	0.70		0.70	0.60	0.66		0.71	0.58	0.65	0.64	0.62	0.72	
JL	0.77	0.74	0.42	0.75	0.60	0.67	0.62	0.75	0.71		0.73	0.79	0.69	0.56	0.68	0.68
RVF	0.60	0.59	0.46	0.60	0.44	0.65	0.56	0.70	0.58	0.73		0.72	0.57	0.56	0.71	0.78
JMQ	0.68	0.65	0.54	0.64	0.52	0.69	0.62	0.70	0.65	0.79	0.72		0.65	0.67	0.75	0.75
RG	0.58	0.58	0.45	0.61	0.54	0.55	0.47	0.66	0.64	0.69	0.57	0.65		0.57	0.65	0.62
TA	0.57	0.57	0.57	0.64	0.35	0.63	0.52	0.54	0.62	0.56	0.56	0.67	0.57		0.66	0.61
D20	0.71	0.67	0.63	0.69	0.32	0.75	0.63	0.71	0.72	0.68	0.71	0.75	0.65	0.66		
D100	0.58	0.61	0.36	0.62	0.56	0.74	0.67	0.60		0.68	0.78	0.75	0.62	0.61		
Average	0.59	0.60	0.45	0.62	0.46	0.59	0.51	0.65	0.64	0.69	0.60	0.65	0.58	0.57	0.66	0.63

Source: Authors' calculation based on Wine Services (2015) data. WS: Wine Spectator; RP: Robert Parker; JR: Jancis Robinson; JD: Jacques Dupont; BD: Bettane & Desseauve; NM: Neal Martin; D20: Decanter20; JS: James Suckling; D100: Decanter100; RVF: La Revue du Vin de France; JCL: Jeannie Cho Lee; AG: Antonio Galloni; JL: Jeff Leve; JMQ: Jean-Marc Quarin; TA: Tim Atkin; RG: Rene Gabriel.

Figure 1

Distribution Functions for Each Transformation and Robert Parker's Score Distribution



Source: Authors' calculation based on Wine Services (2015) data.

function of the Robinson's scores from the interval [12,20] into the interval [70,100][4]. The best way to judge the relevance of this transformation is to compare the respective distribution functions. Figure 1 displays the distribution functions of Jancis Robinson's scores after each transformation, compared to Robert Parker's score distribution function.

The distribution of Jancis Robinson's transformed scores is closer to Robert Parker's distribution with the affine function. Still, one might argue that Jancis Robinson's transformed scores are still underrated compared to the grading system of Robert Parker. More than half of Robert Parker's scores are above 90/100, against only 8% for the Robinson's scores computed with the affine function. As a result, a 90/100 for Robert Parker is a much lower evaluation of quality than a 90/100 for Jancis Robinson with the affine function. A satisfactory transformation of the scores should both put the scores on the same scale and convey the same value to each score. Jancis Robinson's transformed scores should then follow the same distribution function as Robert Parker's scores. Such a function exists and is non-parametrically tractable.

The theoretical framework is the following. Posit that quality of Bordeaux wines is a random variable. The experts evaluate this quality along a scale of their choice, according to their preferences and to their utilization of their scales. Let F be the distribution function of Jancis Robinson's scores, and G be the distribution function of

[4] This affine conversion formula of $x \in [12, 20]$ into $y \in [70, 100]$ is $y = \frac{30}{8}x + 25$.

Robert Parker's scores. These functions express both experts' grading scales as well as their respective appreciation of Bordeaux wines. These differences in scales and in overall appreciation of Bordeaux wines tackle the comparison between grades given by two experts. The method controls for both issues at the same time.). Recall that our objective is to utilize Jancis Robinson scores and translate them into Parker scores, accounting for the fact that Jancis Robinson usually awards lower scores.

We apply the function $G^{-1} \circ F$ in order to obtain the same distribution function for the Jancis Robinson transformed scores and Robert Parker raw scores. This uses the following classical property of probability distributions. Let F_X and F_Y be the distribution of the continuous random variables X and Y, then the random variable $F_Y^{-1} \circ F_X(X)$ has the same probability distribution as Y, F_Y^{-1} being the generalized inverse of F_Y. To avoid any selection bias, the two empirical distributions are computed on a common sample, which contains all wines with a score from each of the two experts. For the chosen pair of experts, the sample includes 1,833 observations.

Let s_{ik} be the score given by expert i to wine k, and I_i be the list of the wines graded by expert i. The procedure is the following:

1) For each expert i, we compute the empirical distribution function

$$F_i(x) = \frac{1}{card(I_i)} \sum_{k \in I_i} 1\{s_{ik} \leq x\}$$

2) For any chosen expert j (here we have chosen Parker), we compute the generalized inverse of F_i:

$$F_j^{-1}(y) = \inf \{x \in \mathbb{R} \mid \hat{F}_j(x) \leq y\}$$

3) The conversion function of the grades of expert i into the scale of expert j is given by:

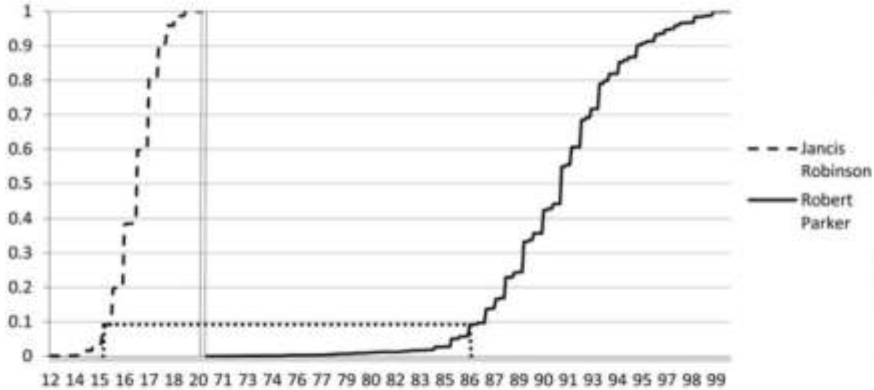
$$\varphi_{ij}(x) = F_j^{-1}\left(\hat{F}_i(x)\right)$$

Figure 2 provides a graphical illustration of our method. As an example, we evaluate the image of a 15/20 from Jancis Robinson on the Robert Parker scale.[5] 15/20 is

[5] The procedure is symmetrical, i.e., it is possible to turn the scores of any expert into the scale of any other expert. Also, it is self-consistent as the conversion function from expert A to expert B is the inverse of the conversion function from expert B to expert A, for all scores observed in the data. For instance, as the data contains a 90/100 from Parker, if we put this score into another expert's scale and turn it back into Parker's

Figure 2

Original Method Using the Empirical Distribution Functions



Source: Authors' calculation based on Wine Services (2015) data.
Note: The double vertical lines stands for the gap on the x-axis between 20 and 70.

the quantile of order 0.092 for Jancis Robinson's distribution function, which means that 9.2% of the Jancis Robinson scores are less than or equal to 15/20. On the Robert Parker distribution function, we read that this quantile is 86/100. We obtain that a 15/20 given by Jancis Robinson is worth a 86/100 given by Robert Parker. In the situation previously stated, this method allows the Jancis Robinson score to be turned into the Robert Parker scale. The average of the two scores is a synthetic indicator of all available information, and can be directly compared with Parker scores if Jancis Robinson scores are missing.

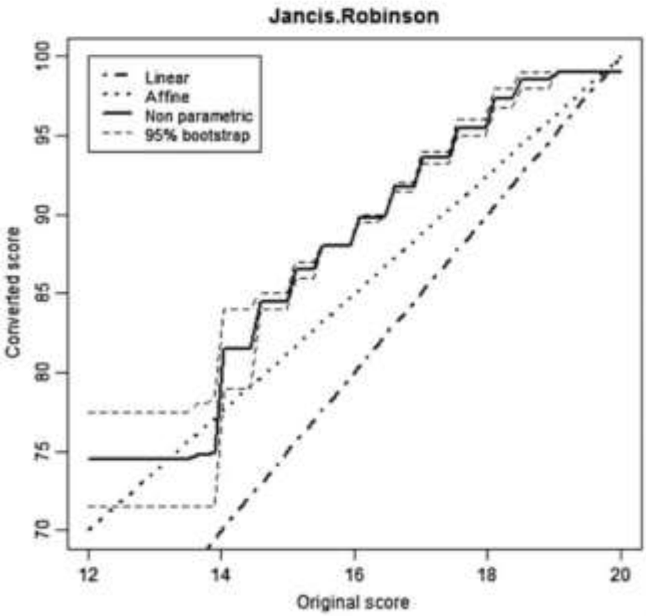
Applying the same method for all existing scores from Jancis Robinson, we obtain a non-parametric function which ensures that the image scores have the same distribution as the Robert Parker scores. Figure 3 compares the plots of three functions, i.e., linear, affine and non-parametric.[6]

scale, we will always end up with a 90/100. This works for all observed scores in the data. However, to be comprehensive, it is not exactly the case with the scores that are unobserved in the data (because the empirical cumulative distribution function is not bijective). The transformation function combined with it generalized inverse does not necessarily give the exact same score. Indeed, the procedure always end up with a score that is originally observed in the data. A simple way to overcome this asymmetry would be to linearly interpolate the empirical distribution function, so as to obtain only bijective functions. As we have not meant the procedure to be applied to scores out of the sample, this is not a major issue for the scope of this paper. Furthermore, considering the large size of our database, the observed scores most likely include all potential scores, so that symmetry is guaranteed for arguably every possible score and for each expert.

[6] The confidence bands have been obtained by bootstrapping the curve 1,000 times. That is to say, we resampled our data 1,000 with replacements, and conducted this procedure for each sample. For each score, we then obtained 1,000 estimates of the converted score. The bootstrap confidence interval is given by the quantiles of order 0.025 and 0.0975 of each score.

Figure 3

Plot of Three Transformation Functions



Source: Authors' calculation based on Wine Services (2015) data.

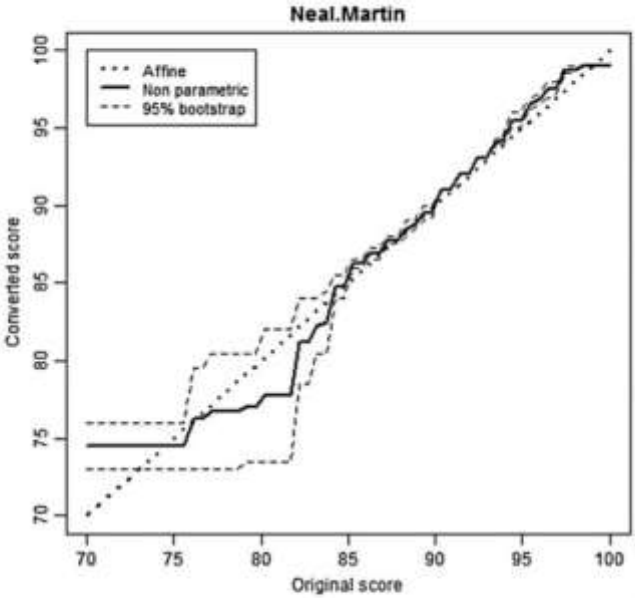
The non-parametric function is irregular on the half-open interval [12,14[. In fact, this interval only concerns 5 observations and 0.4% of the distribution of the Jancis Robinson scores. It corresponds to the half-open interval [70,81.5[for Robert Parker. As a result, the confidence interval is wide below 14/20, so that our conversion is not significantly different from the affine conversion for low grades. However, for high scores, the non-parametric conversion yields significantly higher grades out of 100 than the affine one.

Besides, while the correlation coefficients are neither affected by the linear nor by the affine conversion, the non-parametric method slightly alters the coefficients between the experts. The coefficients computed after conversion are given in Annex 2. The change in the correlation coefficient provides a measure of the non-linearity of the non-parametric conversion. This is measured by the absolute difference between the coefficients before and after conversion in Appendix 2.

This method can also be applied for two experts who both score out of 100. Figure 4 plots the non-parametric function which turns Neal Martin scores into the Robert Parker scale. We find the same regularity issue below 85 points, but the function suggests that Robert Parker has been less reluctant than Neal Martin to grant scores above 95/100. For instance, a 97/100 by Neal Martin is as rare as a 98/100 by Robert Parker. Still, the non-parametric conversion does not represent

Figure 4

Conversion of Neal Martin Scores into Robert Parker Scale



Source: Authors' calculation based on Wine Services (2015) data.

much change compared to the identity function. Our method is more valuable for experts who do not grade on the same scale. All conversion curves are displayed in Appendix 1, along with the affine and the linear ones (which are only different for the experts who grade out of 20). For the latter in particular, the results of the non-parametric method are significantly different from the output of the affine conversion.

IV. Example of Outcomes

Our conversion method facilitates various kinds of comparisons between scores, whether among winemakers, appellations or vintages. We hereafter provide an insight into the possible outcomes. While the general method allows the scores of any expert to be converted into any other expert's scale, we have chosen to convert all scores into the Robert Parker scale. Since he is commonly referred to as the most influential expert for Bordeaux wines (see notably Hadj Ali et al., 2008; Masset et al., 2015), we assume that his scale is the most familiar for the reader.

Table 4 displays all available 2013 *primeurs* scores for a subsample of twenty Bordeaux properties. Columns 2 to 4 reports the average of the available scores transformed by the linear, the affine and the non-parametric function, respectively. Our

Table 4

Raw *primeur* Scores for a Subsample of Vintage 2013 and Mean Scores Computed for the Three Methods

Wine	*Score Linear*	*Score Affine*	*Score Non Parametric*	*sd*	*RP*	*NM*	*JR*	*WS*	*AG*	*BD*	*JD*	*JS*	*JL*	*RVF*	*D*	*JMQ*	*RG*	*TA*
Angelus	89.4	91.7	92.7	1.87	91.5	91	17.5		91	18.75		92.5	92	16.25	90.25	16.5	17	95
Ausone	90.3	92.6	93.7	2.27	94	92	17.5		91	19	17	91.5	94	16	92	16.75	19	94
Cheval Blanc	89.1	91.7	92.6	1.25	90	92	17	92.5	91	18	16.5	93.5	93.5	16.25	92	16.5	18	93
Clinet	86.7	89.7	90.4	1.2	92	91	16	88.5	91.5	16		90.5	92	16	89	15.5	17	91
Eglise Clinet	91.2	93,0	94,0	1.93	93	95	17.5	90.5	91.75	17.5		93.5	95	17.25	91	16.75	19	96
Evangile	88.1	90.8	91.9	1.98	88.5	92	18	91.5	93	17.75	16	90.5	91	16.25	90.25	16	17	92
Gazin	85.6	89.3	90.2	1.5	91	90	16	87.5		15.5	16.75		90.5	16.75	89	15.75	17	91
Grand Vin de Latour	89.1	91.8	92.8	1.55	89		17	91.5	92	17.25	16.75	92.5	92	17.5	94	16	18	95
Haut Brion	90.1	92.2	93.1	1.85	91	90	16.5	92.5	92.5	18	16.75	92.5	93	16.5	94		19	93
La Conseillante	86.7	89.9	90.8	1.7	90	91	15	89.5	91.5	17	15.5	90.5	92	17	90	16	17	92
La Violette	88.6	91.1	91.7	2.49	87	93			88	17.75			93.5	16.75		16.25	18	93
Lafite Rothschild	88.6	91.3	92.4	1.8	88	92	17	90.5	91.5	17.75	16.25	92.5	91	16.25	94	16	18	95
Lafleur	90.6	92.8	93.8	1.84	90	94	18		93	18		93.5	93.5	17.5	93	16.75	17	95
Le Gay	86.5	89.6	90.3	2.64	86	91	15		90.5	17		91.5	93	15	88	16	18	93
Margaux	89.9	92.1	92.9	1.53	89	92	16.5	91.5	92.5	17.25	16.5	94.5	93.5	17.5	94		18	94
Mouton Rothschild	89.6	92,0	92.8	1.79	92	93	17	92.5	89	18.25	17	92.5	93	17.5	94	16.75	17	92
Pavie	88.6	91.3	92.1	1.9	93	92	16		92.25	18		91.5	93.5	15.5	93	16.25	18	90
Petrus	90.2	92.5	93.5	1.82	91.5	91	18.5		91	18	17.5	92.5	94	17	92	16.5	18	94
Trotanoy	87.5	90.7	91.7	1.6	92.5	91	16	91.5	91	17	15	90.5	93.5	16	91	16	18	95
Vieux Chateau Certan	88.0	90.6	91.5	1.93	87.5	93	17.5	91.5	91.5	17.5	15	90.5	91.5		91	16.25	17	92

Source: Authors' calculation based on Wine Services (2015) data.

Notes: sd: standard deviation of the scores obtained from non-parametric method. RP: Robert Partker ; NM: Neal Martin; JR: Jancis Robinson; WS: Wine Spectator; AG: Antonio Galloni; BD: Bettane et Desseauve; D: Decanter; JD: Jacques Dupont; JS: James Suckling; JL: Jeffe Leve; RVF: Revue du Vin de France ; JMQ: Jean-Marc Quarin; RG: Rene Gabriel; TA: Tim Atkin.

Table 5
Mean Vintage Score for Robert Parker and Jancis Robinson with and without Transformation

Vintage	*Number of Observations*	*Robert Parker*	*Jancis Robinson - Non-Parametric Function*	*Jancis Robinson - Raw Scores*
2003	126	90.5	89.9	16.1
2004	69	91.3	92.0	16.6
2005	174	91.8	91.7	16.6
2006	116	91.6	92.2	16.7
2007	196	88.7	90.5	16.2
2008	198	91.0	91.9	16.6
2009	195	92.6	92.4	16.7
2010	201	92.4	92.4	16.7
2011	186	90.0	91.2	16.4
2012	194	90.5	92.3	16.7
2013	168	88.9	91.1	16.3

Source: Authors' calculation based on Wine Services (2015) data.
Note: We lack Jancis Robinson primeurs scores for vintages 2000, 2001, 2002 and 2014.

non-parametric method yields the highest scores, as it transposes the scores on the scale of Robert Parker, used to giving high scores compared to his peers. Overall, the other experts mitigate the negative opinion of Robert Parker of the 2013 vintage, as the mean score is often above Robert Parker's grade.

The last column of Table 4 provides the standard deviation of the scores for each wine. As our method displays all scores on the same scale, it is now possible to compute the relevant standard deviation for each wine across experts. This provides a measure of judge concordance for each wine: the lower the deviation among the scores, the more reliable is the mean score. Château Clinet shows the highest level of agreement among the raters with a standard deviation of 1.2 while Château Le Gay shows the largest dispersion with a standard deviation of 2.64.

Another possible outcome is to facilitate the comparison between vintages for two experts. Table 5 displays the mean scores of vintages 2003 to 2013 for Robert Parker and Jancis Robinson with and without the transformation of Jancis Robinson's scores. Expressing the two assessments on one scale makes them comparable. Our transformation highlights that Jancis Robinson was much more lenient with the 2007 and 2013 vintages than Robert Parker, and that she apparently enjoyed vintage 2012.

V. Conclusion

This paper employs a simple methodology to express the scores of various wine experts on the same rating scale. It facilitates the comparability of the scores among experts and allows to calculate an average of all available wine scores.

Nevertheless, several issues still have to be addressed. Who has to be the expert of reference? Robert Parker seems to be the natural candidate but he has now retired and stopped tasting the Bordeaux *en primeur* in 2015. How to interpret the standard deviation in the cases where wines are not tasted by the same number of experts? Does a standard devation calculated on the basis of 2 scores provide the same information as a standard devation calculated on the basis of 15 scores in terms of consensus? Other questions will certainly have to be addressed and we hope that this paper will induce further research to improve our methodology.

References

Ali, H.H., and Nauges, C. (2007). The pricing of experience goods: the example of *en primeur* wine. *American Journal of Agricultural Economics*, 89(1), 91–103.

Ali, H.H., Lecocq, S., and Visser, M. (2008). The impact of gurus: Parker grades and *en primeur* wine prices. *The Economic Journal*, 118(529), F158–F173.

Ashenfelter, O. (2008). Predicting the quality and prices of Bordeaux wine. *The Economic Journal*, 118(529), F174–F184.

Ashenfelter, O., Ashmore, D., and Lalonde, R. (1995). Bordeaux wine vintage quality and the weather. *Chance*, 8(4), 7–14.

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

Ashton, R.H. (2014). Wine as an experience good: Price versus enjoyment in blind tastings of expensive and inexpensive wines. *Journal of Wine Economics*, 9(2), 171–182.

Bodington, J.C. (2015). Evaluating wine-tasting results and randomness with a mixture of rank preference models. *Journal of Wine Economics*, 10(1), 31–46.

Cardebat, J.M., Figuet, J.M., and Paroissien, E. (2014). Expert opinion and Bordeaux wine prices: An attempt to correct biases in subjective judgments. *Journal of Wine Economics*, 9(3), 282–303.

Mahenc, P., and Meunier, V. (2006). Early sales of Bordeaux grands crus. *Journal of Wine Economics*, 1(1), 57–74.

Hodgson, R.T. (2008). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.

Masset, P., Weisskopf, J.P., and Cossutta, M. (2015). Wine tasters, ratings, and *en primeur* prices. *Journal of Wine Economics*, 10(1), 75–107.

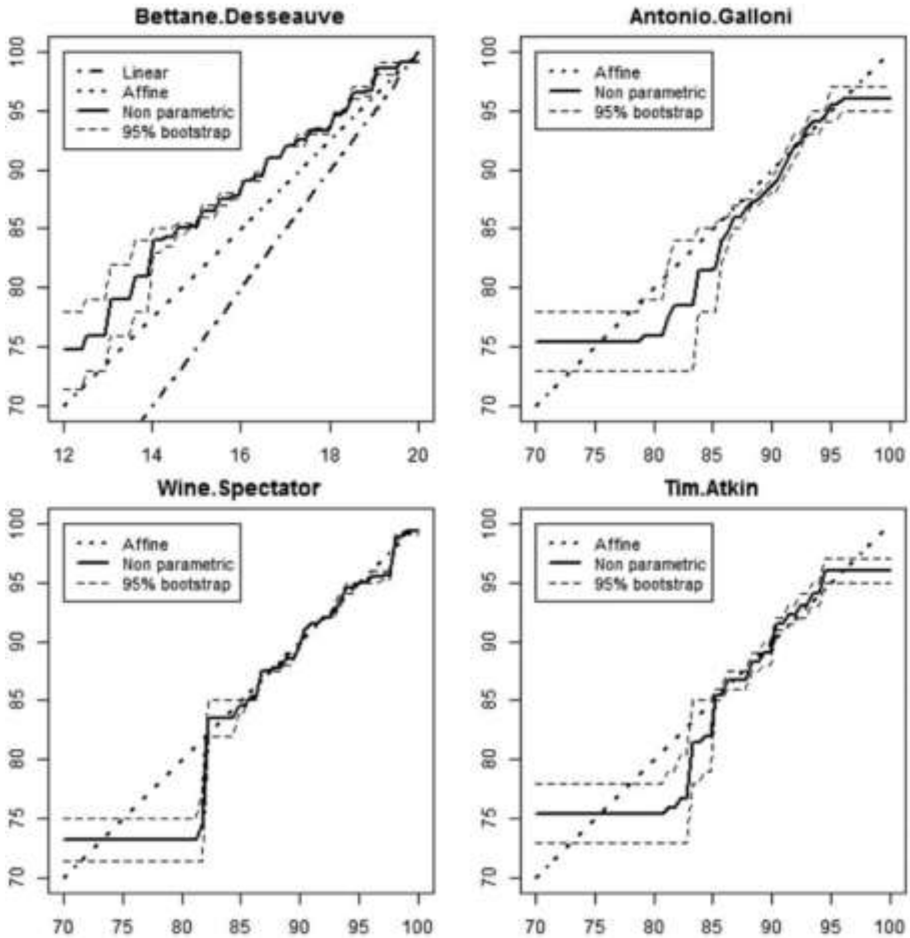
Olkin, I., Lou, Y., Stokes, L., and Cao, J. (2015). Analyses of wine-tasting data: A tutorial. *Journal of Wine Economics*, 10(1), 4–30.

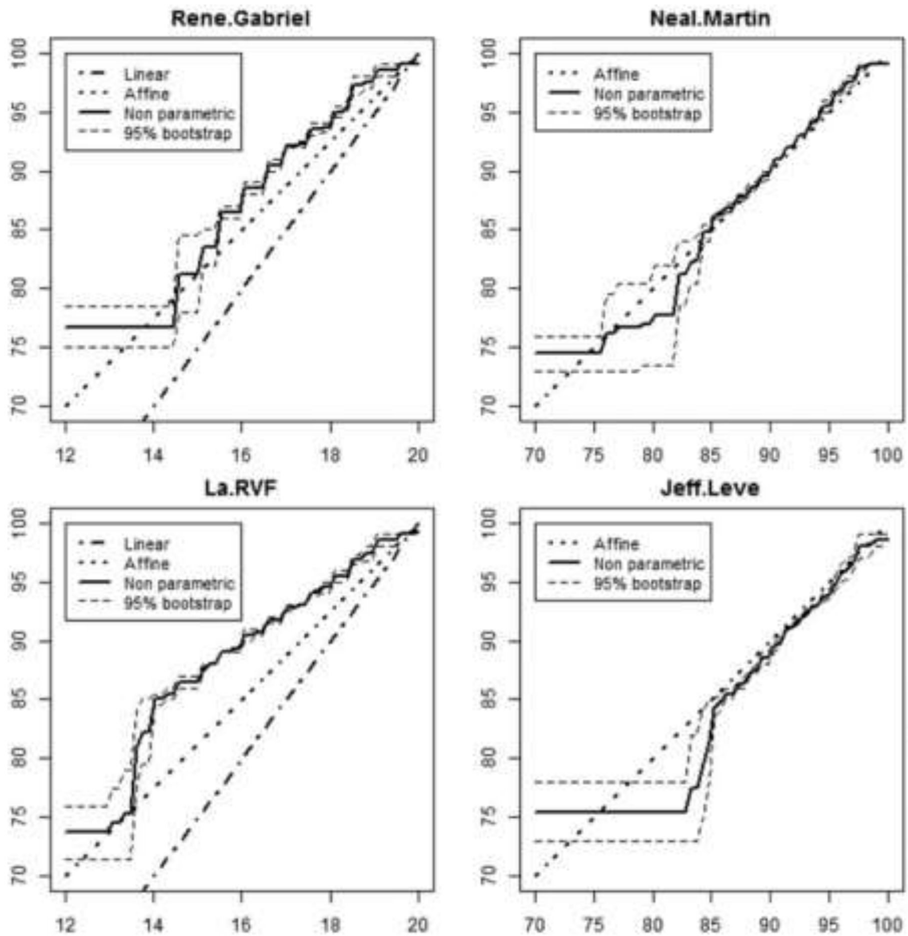
Quandt, R.E. (2007). On wine bullshit: some new software? *Journal of Wine Economics*, 2(2), 129–135.

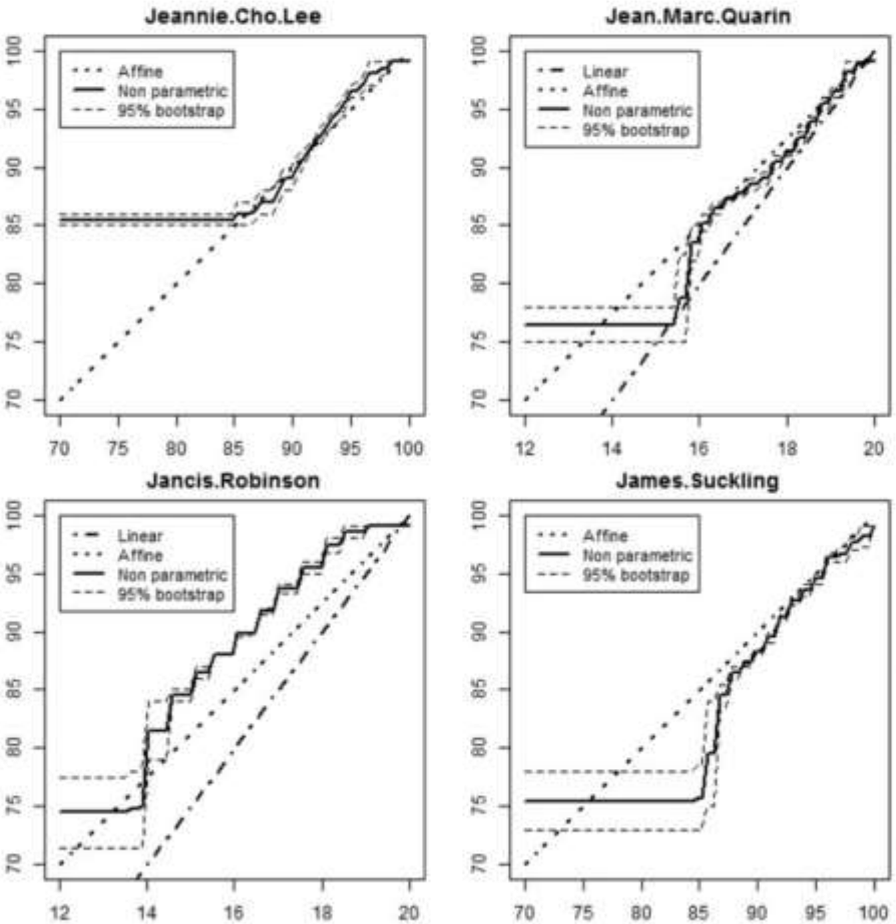
Storchmann, K. (2012). Wine economics. *Journal of Wine Economics*, 7(1), 1–33.

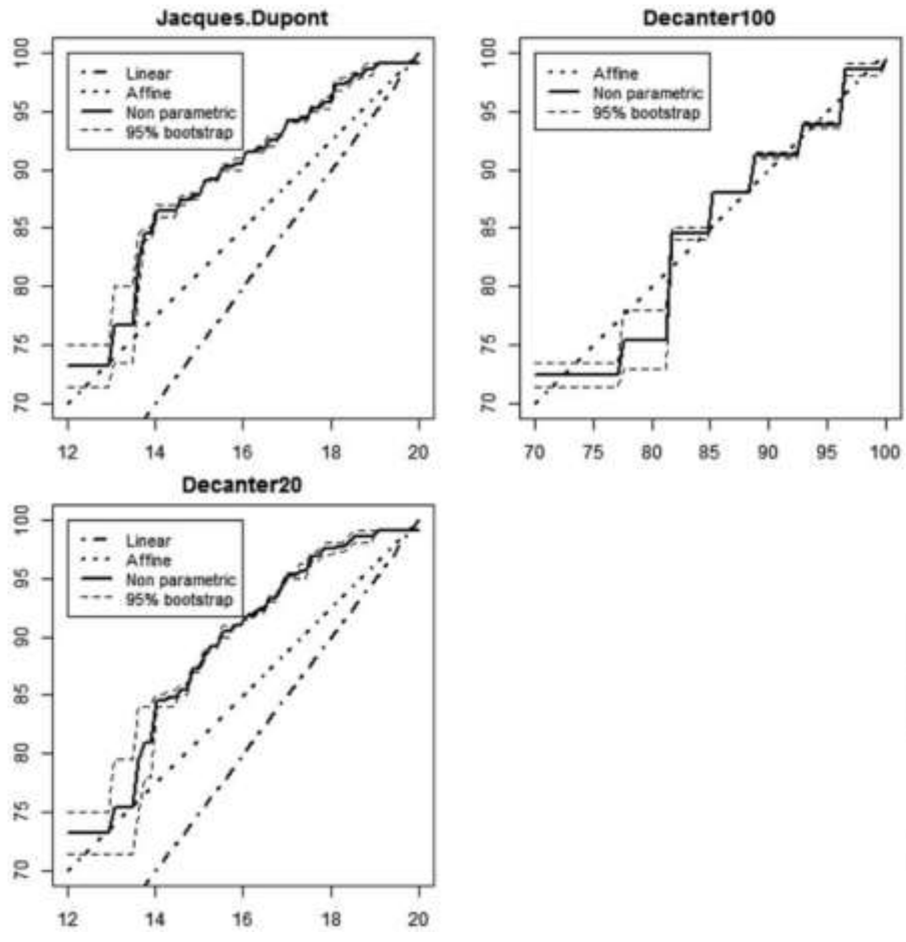
Wine Services (2015). http://www.wine-services.com.

Appendix 1
Conversion Functions into Parker's Scale




Rene.Gabriel
Linear
Affine
Non parametric
95% bootstrap
Neal.Martin
Affine
Non parametric
95% bootstrap
La.RVF
Linear
Affine
Non parametric
95% bootstrap
Jeff.Leve
Affine
Non parametric
95% bootstrap


Jeannie.Cho.Lee
Affine
Non parametric
95% bootstrap
Jean.Marc.Quarin
Linear
Affine
Non parametric
95% bootstrap
Jancis.Robinson
Linear
Affine
Non parametric
95% bootstrap
James.Suckling
Affine
Non parametric
95% bootstrap


Jacques.Dupont
Linear
Affine
Non parametric
95% bootstrap
100
95
90
85
80
75
70
12
14
16
18
20
Decanter100
Affine
Non parametric
95% bootstrap
100
95
90
85
80
75
70
70
75
80
85
90
95
100
Decanter20
Linear
Affine
Non parametric
95% bootstrap
100
95
90
85
80
75
70
12
14
16
18
20

Appendix 2
Correlation Matrix after Conversion

	RP	*NM*	*JR*	*WS*	*AG*	*BD*	*JD*	*JS*	*JC*	*JL*	*RVF*	*JMQ*	*RG*	*TA*	*D20*	*D100*
RP		0.58	0.42	0.60	0.39	0.59	0.49	0.67	0.68	0.77	0.58	0.67	0.58	0.59	0.70	0.55
NM	0.58		0.49	0.60	0.57	0.57	0.49	0.67	0.58	0.74	0.57	0.65	0.57	0.59	0.66	0.59
JR	0.42	0.49		0.48	0.17	0.49	0.36	0.44	0.55	0.40	0.43	0.50	0.45	0.58	0.61	0.36
WS	0.60	0.60	0.48		0.54	0.59	0.46	0.68	0.70	0.69	0.56	0.62	0.60	0.65	0.67	0.55
AG	0.39	0.57	0.17	0.54		0.44	0.33	0.47		0.57	0.42	0.47	0.54	0.35	0.18	0.58
BD	0.59	0.57	0.49	0.59	0.44		0.48	0.62	0.71	0.65	0.62	0.66	0.55	0.64	0.74	0.70
JD	0.49	0.49	0.36	0.46	0.33	0.48		0.54	0.56	0.59	0.52	0.56	0.46	0.53	0.60	0.62
JS	0.67	0.67	0.44	0.68	0.47	0.62	0.54		0.66	0.72	0.64	0.65	0.63	0.54	0.67	0.55
JC	0.68	0.58	0.55	0.70		0.71	0.56	0.66		0.70	0.57	0.62	0.64	0.63	0.72	
JL	0.77	0.74	0.40	0.69	0.57	0.65	0.59	0.72	0.70		0.69	0.75	0.66	0.57	0.67	0.62
RVF	0.58	0.57	0.43	0.56	0.42	0.62	0.52	0.64	0.57	0.69		0.69	0.54	0.54	0.66	0.69
JMQ	0.67	0.65	0.50	0.62	0.47	0.66	0.56	0.65	0.62	0.75	0.69		0.63	0.67	0.71	0.68
RG	0.58	0.57	0.45	0.60	0.54	0.55	0.46	0.63	0.64	0.66	0.54	0.63		0.58	0.65	0.63
TA	0.59	0.59	0.58	0.65	0.35	0.64	0.53	0.54	0.63	0.57	0.54	0.67	0.58		0.66	0.62
D20	0.70	0.66	0.61	0.67	0.18	0.74	0.60	0.67	0.72	0.67	0.66	0.71	0.65	0.66		
D100	0.55	0.59	0.36	0.55	0.58	0.70	0.62	0.55		0.62	0.69	0.68	0.63	0.62		
Average	0.59	0.59	0.45	0.60	0.43	0.60	0.51	0.61	0.64	0.65	0.58	0.63	0.58	0.58	0.64	0.60
Absolute Difference	0.000	0.002	0.011	0.008	0.018	0.003	0.010	0.013	0.004	0.004	0.021	0.011	0.007	0.016	0.005	0.025

Source: Authors' calculation based on Wine Services (2015) data. WS: Wine Spectator; RP: Robert Parker; JR: Jancis Robinson; JD: Jacques Dupont; BD: Bettane & Desseauve; NM: Neal Martin; D20: Decanter20; JS: James Suckling; D100: Decanter100; RVF: La Revue du Vin de France; JCL: Jeannie Cho Lee; AG: Antonio Galloni; JL: Jeff Leve; JMQ: Jean-Marc Quarin; TA: Tim Atkin; RG: Rene Gabriel.