Journal of Wine Economics, Volume 6, Number 2, 2011, Pages 160–178

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Improving Experts' Wine Quality Judgments: Two Heads Are Better than One*

Robert H. Ashton[a]

Abstract

This paper reports an analysis of data from the "Judgment of Paris," the 1976 blind tasting of California and French wines that revolutionized the wine world. Using both empirical and analytical methods, I demonstrate that the wine quality judgments of the renowned experts who participated in the Paris tasting would have been improved simply by averaging the quality ratings of two or more of the judges. Moreover, I explore both *how many* of the Paris judges should be included in the average and *which ones* they should be. The results have implications for the practical issue of choosing judges to include in tasting panels that award prizes or provide expert advice to consumers, as well as for better understanding the variability in the price-quality association across hedonic wine pricing studies. (JEL Classification: C93)

I. Introduction

Expert judgment is extremely influential in many areas of life – from medical diagnosis to judicial interpretation to investment advice to grant proposal evaluation to weather forecasting to choosing which wines to purchase and consume. In some areas, the judgments of individual experts "move the market," e.g., the judgments of Warren Buffett in the corporate world or Robert Parker in the wine world. Unfortunately, the high regard in which we hold individual experts (and perhaps our dependence on them) can blind us to the fact that the judgments of even the *best* experts can be improved, often by means that at first may seem too simple to be credible.

A particularly effective way to improve expert judgment, and one that rests on a rich foundation of theoretical and empirical support, is to combine the judgments

* I am indebted to Jack Soll for conversations that have improved this paper and to Zhenhua Chen for excellent research assistance. I would also like to thank the editor (Karl Storchmann) and the anonymous reviewer for helpful comments.

[a] Fuqua School of Business, Duke University, 100 Fuqua Drive, Box 90120, Durham, NC 27708, email: robert.ashton@duke.edu

© The American Association of Wine Economists, 2011

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

of two or more individual experts and use this "composite judgment" instead of those of the experts themselves. In judgment contexts where the truth – or the "correct answer" – is known, research in many fields has firmly established that composite judgments are almost always more accurate than those of the experts on which they are based.

In this paper I demonstrate the power of composites in the context of the "Judgment of Paris," the May 24, 1976, blind tasting of California and French wines that revolutionized the wine world (Taber, 2006). In that tasting, 11 wine experts (nine renowned members of the French wine establishment plus an Englishman and an American who jointly organized the event) tasted ten red wines (six California and four French) and ten white wines (again, six California and four French), rating each on a 20-point scale. I do not address here the issue of whether California or France "won" the tasting, but focus instead on how the judges' ratings could have been improved (i.e., made more accurate) simply by combining the individual ratings of two or more of the judges.[1]

To allow statements to be made about judgment accuracy, I designate the ratings of one judge (Christian Millau, for reasons described later) as "correct" for purposes of the analysis. I then employ empirical and analytical approaches to examine two questions: (1) *How many* of the Paris judges should have been included in a composite to achieve substantial improvement in accuracy vis-à-vis that of the individual judges? (2) *Which specific judges* should have been included?

The paper proceeds as follows. Section II presents the rationale, and the evidence, for the robust finding that composite judgments, i.e., simple averages of experts' individual judgments, are almost always more accurate than the experts themselves, and it explains why this issue is of particular interest in the context of expert wine quality judgments. Section III describes three characteristics of individual judges that determine the *extent* to which composites are more accurate than individuals: the accuracy of the individual judges and the level of agreement among their judgments (both of which are "given" by the results of the Paris tasting) and the number of judges included in the composite. The impact on composite accuracy of the number of judges included in the composite is explored in two ways – empirically (by forming all possible composites of all numbers of the Paris judges) and analytically (by relying on earlier, more general, theoretical analyses of the power of composites). The theoretical analyses allow me to develop empirical estimates of both *how many* of the Paris judges should be included in a composite and *which ones* they should be. Section IV presents a brief summary and conclusion.

[1] The judges' ratings (especially of the reds) have been statistically analyzed from various perspectives by other researchers (e.g., Ashenfelter and Quandt,1999, Cicchetti, 2004a, 2006; Lindley, 2006; Quandt, 2006, 2007; Hulkower, 2009), typically with the issue of "who won" in mind.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

II. The Power of Composites

It has long been known that the accuracy of composite judgments, formed by simply taking the mean of $k \geq 2$ individuals' judgments, must be equal to or greater than the mean accuracy of the individuals. Using the Pearson correlation between a series of judgments and the corresponding criterion values as a measure of judgment accuracy, Dawes (1970), generalizing earlier work on psychological measurement by Ghiselli (1964, chapter 7), proved that $\rho_{y\bar{x}} \geq \bar{\rho}_{yx}$, where $\rho_{y\bar{x}}$ is the accuracy of a composite of k individuals, expressed as the correlation between $\bar{x}\left(= 1/k \sum_{j=1}^{k} x_j\right)$, or the mean of the $(j = 1, \ldots, k)$ individual judgments for each observation, and y, or the corresponding criterion values, and $\bar{\rho}_{yx}$ is the mean accuracy of the k individuals, i.e., $1/k \sum_{j=1}^{k} \rho_{yx_j}$. Moreover, this inequality between the accuracy of mean judgments and the mean accuracy of individual judgments holds not only for a correlational error metric, but for any error metric involving a convex loss function, e.g., mean absolute error or mean squared error (Larrick and Soll, 2006).[2]

The basic principle can be illustrated as follows: Imagine that (1) two individuals are asked to estimate the number of marbles in a clear jar that sits in front of them, (2) the actual (correct) number of marbles in the jar is 500, and (3) mean absolute error (deviation from 500) is used to compare the error of their mean estimate with the mean of their two individual errors. The relative sizes of their mean (or composite) error vis-à-vis the mean of their individual errors depends on whether their two estimates "bracket" the correct answer and, if so, the width of the bracket (Larrick and Soll, 2006).

Consider first the case where the two individual estimates do not bracket the correct answer: for example, if the two estimates are 520 and 550, the mean of the individual errors is 35 [i.e., the mean of 20 (520–500) and 50 (550–500)] and the composite error is also 35 [(520 + 550)/2–500]. Now assume the two individual estimates are 520 and 450: the mean of the two individual errors is still 35 [mean of 20 (520–500) and 50 (450–500], but the composite error drops to 15 [(520 + 450)/2–500]. Wider brackets around the correct answer favor composites even more strongly; for example, with individual estimates of 620 and 350, the mean of the two individual errors is 135 [mean of 120 (620–500) and 150 (350–500)], while the composite error remains at 15 [(620 + 350)/2–500]. Thus, the no-bracketing case produces equivalent errors, while the bracketing case produces smaller errors for composite estimates. Putting the bracketing and no-bracketing cases together, it is clear that over a series of such estimates using the mean of the two independent

[2] While Dawes (1970) and Ghiselli (1964) were among the first to recognize the power of composites for the field of psychological measurement, the basic insight was developed earlier. For example, the Danish mathematician Johan Jensen proved an analogous inequality for convex functions in 1906 (see, e.g., Rudin, 1987).

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

estimates will result in smaller errors provided there is at least one instance of bracketing.[3]

The superior accuracy of composite judgments has been extensively studied both theoretically and empirically. Key theoretical analyses include Bates and Granger (1969), Hogarth (1978), Clemen and Winkler (1985), and Wallsten, Budescu, Erev, and Diederich (1997). The power of composites has been demonstrated empirically in numerous applied settings including meteorology (e.g., temperature and precipitation forecasts), business (e.g., predictions of corporate earnings, sales volumes, stock prices, and plant and equipment investments), economics (e.g., forecasts of exchange rates, inflation, and unemployment), and in many other settings involving phenomena as varied as population size, property values, football game outcomes, the demand for electricity, tourism activity, and livestock prices.[4]

The present analysis goes beyond the well-established result that composite judgments are more accurate than individual judgments. In this paper I consider the more exacting research questions of (1) the *number* of judgments that should be included in the composite and (2) *which* of a set of available judges should be included. While some empirical research exists on the number of judges to include, very little research exists on which specific judges (of those available) they should be. On the first question, the evidence suggests that including only a small number of judges (say, two to four) in the composite is often sufficient (e.g., Libby and Blashfield, 1978; Makridakis and Winker, 1983; Winkler and Makridakis, 1983; Ashton and Ashton, 1985; see also Clemen, 1989, and Armstrong, 2001). Improvements in accuracy from the inclusion of additional judges decrease rapidly thereafter. On the second question, the scarce theoretical and empirical research that exists provides useful guidance about how one should choose which particular judges to include in the composite (e.g., Hogarth, 1978; Ashton, 1986).

In the specific context of experts' judgments of wine quality, the sensitivity of composite accuracy to the number of judges in the composite is important for a number of reasons. One concerns the results of hedonic pricing studies that examine the objective and sensory determinants of wine prices. One price determinant that has been investigated extensively is the wine quality ratings assigned by individual wine critics (e.g., Robert Parker or James Laube) and expert panels. Studies that use individual wine critics' ratings as an overall measure of quality tend to find a strong association with wine prices (DiVittorio and Ginsburgh, 1996; Jones and Storchmann, 2001; Haeger and Storchmann, 2006;

[3] The superior accuracy of composite judgments is easily seen in simple examples such as this – and in the more general derivations such as that of Dawes (1970). In everyday life, however, the situation is very different: Larrick and Soll (2006) provide an informative account of how "the averaging principle" is neither naturally intuitive nor aided by observation of our own or others' judgments.

[4] The empirical literature is voluminous. Excellent reviews covering studies in these and other areas are provided by Ferrell (1985), Flores and White (1988), Clemen (1989), and Armstrong (2001). The latter two reviews are especially comprehensive.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Roberts and Reagans, 2007; Hadj Ali, Lecocq and Visser, 2008; Gibbs, Tapia and Warzynski, 2009).[5] On the other hand, studies that use quality ratings provided by *panels* of judges sometimes find a strong association between quality and price (Oczkowski, 1994; Benjamin and Podolny, 1999; Schamel and Anderson, 2003; Schnabel and Storchmann, 2010), but more often find, at best, a weak association (Combris, Lecocq and Visser, 2000; Oczkowski, 2001; Lecocq and Visser, 2006; Troncosco and Aguirre, 2006; San Martin, Troncosco and Brümmer, 2008).

There are at least two explanations for the conflicting results concerning the strength of the price-quality association in such studies. One concerns the amount of information about wine quality that is available to consumers. Tellis and Wernerfeldt's (1987) reanalysis of nine empirical studies involving more than 1,200 products found tremendous variability in the correlation between price and quality: while the mean (median) correlation was .27 (.31), the range was −.26 to 1.00. Products for which greater amounts of information about quality was available to consumers showed substantially higher levels of association between price and quality. Thus, one explanation for conflicting results concerning the price-quality association across hedonic wine pricing studies is differential availability of information about the particular wines studied.

Another explanation, and the one considered here, is differential measurement error across the various studies. Lecocq and Visser (2006) observe that a possible explanation for the weak association between wine prices and wine quality ratings in studies involving panels of judges is measurement error in the quality ratings assigned by the individual members of the judge panels.[6] To address this problem, Lecocq and Visser (2006, p. 50) suggest that "augmenting the number of members in a jury might reduce the variance of the mean of judgment errors, thereby ... [increasing] the correlation coefficient between price and jury grade ...." As demonstrated by the empirical literature on composite judgments (and confirmed here with the Paris judges), both the mean and variance of judgment errors are strongly influenced by the number of judges included in the composite.

Another reason this issue is important is the practical necessity of deciding how many judges to include in tasting panels that award medals at wine competitions (or that simply provide expert advice to consumers). Many such panels involve judges who independently assess wines using a numerical rating system, followed by a statistical combination (typically the simple average) of their independent ratings. The number of judges varies widely across such panels. Gawel and Godden (2008), for example, combine the independent ratings of three judges, as is

[5] The results of Haeger and Storchmann (2006), however, demonstrate the importance of controlling for other, more fundamental, indicators of wine quality. While these authors find a significant association between quality ratings and price (table 6), the effect is substantially weaker after controlling for producer and certain climate variables (table 9).

[6] Unwin (1999) and Oczkowski (2001) also consider the issue of measurement error in hedonic wine pricing studies, especially measurement error in the quality variable.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

commonly done in the Australian wine show system. Hodgson (2008) reports a study of the California State Fair Wine Competition in which the independent ratings of four judges were combined into overall panel assessments. Large-scale tastings conducted in France by the *Institut National de la Consommation* involved four independent tasters for Bordeaux wines (Combris, Lecocq and Visser, 1997) and five independent tasters for Burgundies (Combris, Lecocq and Visser, 2000). As Goldwyn and Lawless (1991) explain, the procedures used by the Beverage Testing Institute in the American Wine Competition typically involve panels of five independent tasters (actually, seven, but with high and low scores discarded). Lawless, Liu and Goldwyn (1997) report research involving several different panels whose membership includes six to eight independent judges. If one held all else constant in studies like these, the impact on the panel results of the varying numbers of independent judges would, I believe, be substantially greater than intuition might suggest.

III. Analysis and Results

The analyses require that a "correct" rating be specified for each wine. Obviously, correct ratings that are objectively determined and independent of the tasters do not exist. It is often maintained, however, that one judge in a group can usefully be designated as the correct judge, e.g., for the purpose of training other judges to mimic the designated judge (Cicchetti, 2004b, 2006), and this approach is taken here. As Cicchetti (2004a, p. 217) suggests, the designated correct judge ideally would have "a more stringent and well-defined set of scoring criteria" than those not so designated. While it is impossible to know the scoring criteria of the Paris judges, there are features of Christian Millau's ratings that make him a good candidate for designation as the "correct" judge. For example, for the red wines, only two of the judges assign lower average ratings across the 10 wines, suggesting Millau is a "hard grader" for reds, while his average rating for the white wines is at the median of all judges. For both reds and whites, the variance of his ratings (5th largest for reds; 4th largest for whites) suggests he is among the more discriminating judges. Thus, to demonstrate the power of composite judgments, the ratings of Christian Millau are assumed to be correct in all of the analyses.[7]

The extent to which composites of judges are more accurate than the judges themselves depends, in part, on both the accuracy of the individual judges and the extent to which the judges agree in their ratings. The impact on composite accuracy of both individual accuracy and inter-judge agreement is addressed in detail later in the paper. At this point, I begin by presenting descriptive statistics on these two

[7] The *specific* results reported below depend, of course, on the designation of Millau as the "correct" judge. However, the *general* result (that composite judgments are more accurate than individual judgments in this setting) does not. To understand why, see the discussion of Equations (1) and (2) later in the paper.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Table 1
Individual Accuracy

Judge	*Red Wines*	*White Wines*
1 Pierre Bréjoux	.437	.308
2 Aubert de Villaine	.862	.829
3 Michel Dovaz	.072	.155
4 Patricia Gallagher	−.119	.183
5 Odette Kahn	.716	.629
6 Christian Millau[a]	_[a]	_[a]
7 Raymond Oliver	.499	.704
8 Steven Spurrier	.195	−.211
9 Pierre Tari	−.471	.531
10 Christian Vannequé	.341	.462
11 Jean-Claude Vrinat	.763	.707
Mean	.329	.430

[a]The ratings of Judge 6 are used as the criterion, i.e., the "correct" ratings.

determinants of the composite accuracy of the Paris judges. The accuracy of each judge, measured as the Pearson correlation between the judge's ratings and those of Christian Millau, is shown in Table 1. While mean accuracy is somewhat greater for the whites (.430) than for the reds (.329), the range in accuracy across judges is quite striking in both cases – ranging from .829 to −.211 for the whites and from .862 to −.471 for the reds.

The range of agreement (or disagreement) among the judges is even greater. Agreement, measured as the Pearson correlation between the ratings of all pairs of judges, is shown for the red (white) wines in Table 2 (3). For the reds, the range of agreement is .965 to −.616, with 17 judge pairs exhibiting negative correlations, while the range for the whites is .998 to −.069 with three pairs negative.[8] The mean level of pairwise agreement between each judge and the other nine judges is shown at the bottom of Tables 2 and 3, and ranges from .267 to −.024 for the reds and from .570 to .247 for the whites. The grand mean, across all 45 pairs of judges, is .438 for the whites (median = .484) but only .156 for the reds (median = .165).[9] Finally, examination of Tables 2 and 3 reveals that the same judge pairs who agree

[8] Lindley (2006) points out that Judges 7 and 11 appear to have collaborated in their ratings of the whites. Since it is these two judges who produce the correlation of .998, a better indication of the highest pairwise agreement is .941.

[9] The mean correlations with the other judges shown in Tables 2 and 3 are substantially smaller than the correlations with the other judges reported in Quandt (2006). The reason is that Quandt first calculates, for each wine, the mean rating across all of the other judges, and then correlates these mean ratings with those of the remaining judge, whereas I first calculate pairwise correlations and then average them for each pair in which a judge appears. Quandt's aggregation method "averages away" much of the disagreement among the judges, resulting in higher measures of agreement. Quandt also converts the judges' raw scores to ranks before computing correlations, whereas I use the raw scores. When I use Quandt's methodology on the raw scores, the resulting correlations are comparable to those he reports.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Table 2
Pairwise Agreement: Red Wines[a]

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*Judge 7*	*Judge 8*	*Judge 9*	*Judge 10*	*Judge 11*
Judge 1	–									
Judge 2	.306	–								
Judge 3	.313	−.130	–							
Judge 4	.317	−.303	.575	–						
Judge 5	.055	.890	−.307	−.466	–					
Judge 7	−.057	.655	−.616	−.347	.729	–				
Judge 8	.107	−.013	.382	.773	−.089	.166	–			
Judge 9	.086	−.279	.097	.478	−.461	−.251	.206	–		
Judge 10	.949	.211	.266	.289	.012	−.023	.165	.169	–	
Judge 11	.324	.965	−.246	−.326	.891	.710	−.088	−.262	.211	–
Mean	.267	.256	.037	.110	.139	.107	.179	−.024	.250	.242

[a]Judge 6 is omitted because his ratings are used as the criterion, i.e., the "correct" ratings.

in their ratings of the reds often disagree in their ratings of the whites (and vice versa). Thus, there is substantial disagreement among the judges, especially for the reds, and the *patterns* of disagreement across judges vary widely.

A. *Number of Judges to Include*

The third factor (in addition to individual accuracy and inter-judge agreement) that determines the extent to which composite judgments are more accurate than individual judgments is the number of judges ($k = 2, \ldots . n$) included in the composite. The impact of k on the accuracy of composite judgments was evaluated for the Paris judges by forming all possible composites of the ratings of k judges ($k = 2, \ldots ., 10$) and calculating, for each composite, the Pearson correlation between the composite judgments and Millau's ratings. As an example, consider composite judgments based on $k = 3$ judges. There are 120 such composites (i.e., $n!$ / $(n - k)!$ $k! = 10!$ / 7! 3! = 120). For the first set of three judges, composite judgments were formed for each of the red wines (and separately for each of the white wines) by simply averaging the three individual judges' ratings, and the correlation of this 3-judge composite with Millau was calculated.[10] This process was repeated for each of the other 119 3-judge composites, as well as for all composites of all other levels of k judges.[11]

[10] Taking a simple average of the three individual judges' ratings means, of course, that their judgments are weighted equally in forming the composite. The effectiveness of equal weighting vis-à-vis various differential weighting schemes is well established – see, for example, the reviews by Clemen (1989) and Armstrong (2001), as well as the analyses of Dawes and Corrigan (1974), Einhorn and Hogarth (1975), Einhorn, Hogarth and Klempner (1977), and Hogarth and Karelaia (2007).

[11] This entire analysis was repeated using a noncorrelational error metric – the mean absolute percentage error between the "correct" ratings and those of the composites. The results, shown in the Appendix, agree strongly with those presented below.

Table 3
Pairwise Agreement: White Wines[a]

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*Judge 7*	*Judge 8*	*Judge 9*	*Judge 10*	*Judge 11*
Judge 1	–									
Judge 2	.217	–								
Judge 3	.648	−.065	–							
Judge 4	.704	−.027	.484	–						
Judge 5	.605	.793	.187	.104	–					
Judge 7	.545	.849	.142	.147	.933	–				
Judge 8	.541	−.069	.178	.443	.087	.174	–			
Judge 9	.499	.322	.681	.326	.326	.516	.201	–		
Judge 10	.812	.451	.281	.624	.666	.624	.508	.260	–	
Judge 11	.561	.854	.161	.167	.941	.998	.164	.512	.641	–
Mean	.570	.369	.300	.330	.516	.548	.247	.405	.541	.555

[a]Judge 6 is omitted because his ratings are used as the criterion, i.e., the "correct" ratings.

Table 4
Descriptive Statistics for CORR: Red Wines

Number of Judges (k)	*Number of Composites of Size k*	*Mean*	*Variance*	*Low*	*Median*	*High*
1	–	.329	0.1765	−.471	.389	.862
2	45	.504	0.0994	−.316	.638	.852
3	120	.583	0.0589	−.194	.673	.875
4	210	.636	0.0383	−.077	.697	.882
5	252	.673	0.0252	.091	.720	.876
6	210	.700	0.0165	.237	.737	.881
7	120	.721	0.0105	.328	.746	.874
8	45	.738	0.0061	.490	.751	.854
9	10	.752	0.0029	.653	.767	.809
10	1	.764	–	–	–	–

The results are presented in Tables 4 and 5. The "Mean" column of each table represents the mean correlation (CORR) over all composites of k judges. Table 4 (red wines) indicates that the mean CORR of .329 for the 10 individual judges is improved substantially by forming composites. The largest possible composite, consisting of all 10 judges, has a CORR of .764. Forming composites of only two judges is seen to be very effective, as mean CORR jumps from .329 at $k = 1$ to .504 at $k = 2$. Focusing on the *maximum possible* improvement as k increases from 1 to 10 (i.e., .435, or .764 − .329), the average 2-judge composite achieves 40% of this maximum improvement and the average 3-judge composite (CORR = .583) achieves 58% of it. While mean CORR continues to increase as k becomes larger, it increases at a decreasing rate, indicating that the marginal impact of adding judges to the composite decreases as the number of judges increases. Clearly, composites are very effective in improving the accuracy of the individual judges.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Table 5
Descriptive Statistics for CORR: White Wines

Number of Judges (k)	*Number of Composites of Size k*	*Mean*	*Variance*	*Low*	*Median*	*High*
1	–	.430	0.1028	−.211	.496	.830
2	45	.497	0.0444	−.013	.525	.852
3	120	.529	0.0261	.083	.540	.825
4	210	.548	0.0169	.177	.562	.800
5	252	.561	0.0114	.251	.573	.783
6	210	.569	0.0077	.323	.573	.762
7	120	.576	0.0050	.400	.580	.731
8	45	.581	0.0030	.461	.579	.700
9	10	.585	0.0015	.525	.581	.639
10	1	.588	–	–	–	–

Composites are even more effective in improving the accuracy of the *worst* judge (as well as the worst *composites* of judges): The "Low" column of Table 4, representing the least accurate judge and least accurate composite at each level of k, reveals an improvement from $-.471$ at $k = 1$ to .764 at $k = 10$. In fact, the *worst* composites of seven or more judges are as good as or better than the average individual judge. Moreover, the accuracy of even the *best* individual judge (CORR = .862) is improved slightly by composites of size $k = 3$ to $k = 7$. Finally, the variability of CORR decreases substantially as k increases. This is seen in the "Variance" column of Table 4 as well as in the increasing (decreasing) pattern of the "Low" ("High") column: High CORR and Low CORR converge since both the number of composites and the differences among them are smaller as the upper limit of k is approached. Thus, with higher k it becomes less important to identify the best composites or to include the best judges in the composites identified.

The white wines (Table 5) differ somewhat from the reds. For example, while the mean CORR for the individual judges (.430) is greater than that for the reds (.329), forming composites results in a smaller *absolute* degree of improvement: the accuracy of the single 10-judge composite is only .588 for the whites compared to .764 for the reds. In *relative* terms, however, the extent of improvement exceeds that for the reds: again focusing on the maximum possible improvement as k increases from 1 to 10 (i.e., .158, or .588 − .430), the average 2-judge (3-judge) composite achieves 42% (63%) of this maximum improvement.[12] The overall pattern of improvement in mean CORR as k approaches 10 mirrors that for the reds. Finally, and also similar to the reds, the *worst* composites of eight or more judges are more accurate than the average individual judge, and even the accuracy of the *best* individual judge can be improved – in this case, by a 2-judge composite.

[12] The extent to which the maximum possible improvement is achieved by 2- and 3-judge composites is even greater, for both reds and whites, when mean absolute percentage error is used as the error metric (see the Appendix).

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

The foregoing demonstration relies on a brute-force approach – computing the accuracy of *all possible* composites of *all numbers* of k individuals, and evaluating the results (mainly) by examining the mean accuracy of all composites of size k. Fortunately, analytical work exists that allows the accuracy of composite judgments to be estimated from some straightforward information about the individuals' judgments. Hogarth (1978) presents a model that estimates composite accuracy as a function of the number of judges, their mean individual accuracy, and the mean intercorrelation among their judgments (also see Ghiselli, 1964, pp. 181–90). The model is

$$\rho_{y\bar{x}} = \mathrm{k}^{1/2}\bar{\rho}_{yx} \div \left[1 + (k-1)\bar{\rho}_{x_i x_j}\right]^{1/2}, \tag{1}$$

where $\rho_{y\bar{x}}$ is the accuracy of a composite of k judges, expressed as the Pearson correlation between the mean of the ($j = 1, \ldots, k$) judges' individual judgments for each observation, and y, or the corresponding criterion values (here, the ratings of Millau); $\bar{\rho}_{yx}$ is the mean accuracy of the individual judges; and $\bar{\rho}_{x_i x_j}$ is the mean intercorrelation among their judgments. Equation (1) reveals that composite accuracy is an increasing function of the number of judges and their mean individual accuracy, and a decreasing function of the mean intercorrelation among their judgments.

For Equation (1) to provide good estimates of composite accuracy, especially for composites involving small k, Hogarth's analysis suggests that mean intercorrelation must not be low (< .3, approximately) and/or mean intercorrelation must be equal to or greater than mean individual accuracy. In a setting involving quarterly predictions of advertising page sales for *Time* magazine, Ashton (1986) found that Hogarth's model provided remarkably good estimates of composite accuracy even though these two conditions held only moderately well. In the present context, both conditions hold for the white wines: mean intercorrelation (.438) is greater than .3 and is (barely) greater than mean individual accuracy (.430). However, neither condition holds for the red wines: mean intercorrelation (.156) is less than .3 and is also less than mean individual accuracy (.329). Thus, according to Hogarth's analysis, Equation (1) should produce better estimates of composite accuracy for the whites than for the reds. Of course, this is ultimately an empirical issue that can only be settled by examining the data.

The results are presented in Table 6, where the actual (computed) mean accuracy of composites of size k is compared to the level of accuracy estimated by Equation (1). Looking first at the white wines, where both of the conditions specified by Hogarth hold, the degree of correspondence between the actual level of accuracy and that estimated by Equation (1) is very high: the *largest* difference between actual and estimated accuracy (at $k = 10$) is only .024. For the red wines, where neither condition holds, the degree of correspondence is somewhat lower, with differences between actual and estimated accuracy ranging from .071 to .096. Even for the reds, however, the estimated improvements in accuracy from increasing the

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Table 6
Actual and Estimated Composite Accuracy

Number of Judges (k)	*Number of Composites of Size k*	*Red Wines*		*White Wines*	
		Actual[a]	*Estimated*[b]	*Actual*[a]	*Estimated*[b]
2	45	.504	.433	.497	.507
3	120	.583	.497	.529	.544
4	210	.636	.543	.548	.565
5	252	.673	.577	.561	.580
6	210	.700	.604	.569	.590
7	120	.721	.626	.576	.597
8	45	.738	.643	.581	.603
9	10	.752	.658	.585	.608
10	1	.764	.671	.588	.612

[a]Mean computed accuracy for composites of size k (from Tables 4 and 5)[b] Estimated accuracy from Equation (1).

size of the composite closely track the actual improvements. Overall, the results suggest that Equation (1) can be quite useful for estimating *how many* judges a composite should include.

B. *Which Judges to Include*

Equation (1) can also be used to suggest *which* of a set of available judges should be added to an existing k-person composite to best improve its accuracy. Intuitively, one might think that the judge with the greatest individual accuracy should be added to the existing composite. However, the criterion should be to add the judge whose contribution will improve *composite accuracy* by the greatest amount. As Equation (1) suggests, this requires taking into account the intercorrelations among the judges' ratings.[13]

Note that when the $(k+1)$st judge is added to an existing composite of k judges, both mean individual accuracy and mean intercorrelation are likely to change. Defining $\rho_{yx_{k+1}}$ as the individual accuracy of the $(k+1)$st judge and $\bar{\rho}'_{x_i x_j}$ as the mean intercorrelation of the larger set of judges, i.e., including the $(k+1)$st judge, Hogarth (1978) derives the following condition that must hold for addition of the $(k+1)$st judge to improve composite accuracy:

$$\rho_{yx_{k+1}} > \rho_{y\bar{x}}\left\{\left[(k+1)+(k+1)k\bar{\rho}'_{x_i x_j}\right]^{1/2}-\left[k+k(k-1)\bar{\rho}_{x_i x_j}\right]^{1/2}\right\}. \qquad (2)$$

Equation (2) implies that composite accuracy will not necessarily be improved simply by adding a judge whose individual accuracy, $\rho_{yx_{k+1}}$, exceeds the accuracy of the existing k-person composite, $\rho_{y\bar{x}}$. Instead, it may also be necessary that mean

[13] Budescu and Yu (2007) and Broomell and Budescu (2009) provide extensive analyses of the importance of judge intercorrelations to the accuracy of a composite.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

intercorrelation be reduced, i.e., that $\bar{\rho}'_{x_i x_j} < \bar{\rho}_{x_i x_j}$. Similarly, Equation (2) implies that composite accuracy can be improved by adding a judge whose individual accuracy is less than that of the existing composite, so long as mean intercorrelation is reduced sufficiently.[14]

The latter implication of Equation (2) is explored in Table 7, which shows the effects of adding a third judge to an existing 2-judge composite. The results presented earlier establish that while, on average, 2-judge composites are substantially more accurate than individual judges, accuracy can be further improved, again on average, by adding a third judge. The question is *which* judge should be added.

The left-most three columns of Table 7 assume that the existing composite consists of the two judges whose composite accuracy places them at the 50th percentile (median) of the 45 2-judge composites: for the red wines these are Judges 1 and 7 (composite accuracy = .638) and for the whites they are Judges 3 and 5 (composite accuracy = .525). For each of the remaining eight judges, Table 7 shows both their individual accuracy (from Table 1) and the accuracy of the 3-judge composite that includes them. It can be seen that for the reds adding any of Judges 2, 5 or 11 to the median composite will improve its accuracy.[15]

It is perhaps not surprising that 2-judge composites that achieve only median accuracy can be improved by adding a third judge. Can similar improvements be achieved for 2-judge composites that are among the *most accurate* of the 45 composites? The answer is yes, as demonstrated in Table 7 for the 2-judge composites whose accuracy places them at the 75th percentile and the 90th percentile of all 2-judge composites. For the reds, the accuracy of the 75th-percentile (90-percentile) composite can be improved by adding Judge 2, 5 or 7 (Judge 2, 8 or 11).[16]

The results for the 75th- and 90th-percentile composites also demonstrate Equation (2)'s implication, noted earlier, that composite accuracy can be improved by adding a $(k+1)$st judge whose individual accuracy is *less than* that of the existing k-person composite. Consider, for example, the addition of Judge 7 to the 75th-percentile composite consisting of Judges 3 and 11. Although Judge 7's individual accuracy (.499) is substantially below that of the Judge 3/11 composite (.760), the accuracy of the new 3-judge composite increases to .818.[17] Even more striking is the

[14] Conversely, Equation (2) can be used to determine whether *removing* one or more judges from an existing composite would improve its accuracy.

[15] Results for the whites, where median accuracy is lower, are similar: adding any of Judges 2, 7, 9, 10 or 11 will improve composite accuracy.

[16] Again, results for the whites are similar: the accuracy of the 75th-percentile (90th-percentile) composite can be improved by adding Judge 2 or 7 (Judge 5, 7, 9 or 11).

[17] The pairwise agreement results in Table 2 indicate that while Judge 7 exhibits substantial agreement with Judge 11 (correlation of .710), he disagrees strongly with Judge 3 (correlation of $-.616$, the lowest of all 45 pairwise correlations).

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Table 7
Effects of Adding a Third Judge to an Existing Two-Judge Composite

Panel A. Red Wines

2-Judge Composite: Judges 1 and 7 Accuracy of 2-Judge Composite: .638[a]			*2-Judge Composite: Judges 3 and 11 Accuracy of 2-Judge Composite: .760*[b]			*2-Judge Composite: Judges 1 and 5 Accuracy of 2-Judge Composite: .811*[c]		
3rd Judge Added	*Individual Accuracy of 3rd Judge*	*Accuracy of 3-Judge Composite*	*3rd Judge Added*	*Individual Accuracy of 3rd Judge*	*Accuracy of 3-Judge Composite*	*3rd Judge Added*	*Individual Accuracy of 3rd Judge*	*Accuracy of 3-Judge Composite*
2	.862	.835	1	.437	.677	2	.862	.859
3	.072	.605	2	.862	.855	3	.072	.801
4	−.119	.485	4	−.119	.545	4	−.119	.765
5	.716	.809	5	.716	.810	7	.499	.809
8	.195	.612	7	.499	.818	8	.195	.831
9	−.471	.419	8	.195	.689	9	−.471	.715
10	.341	.495	9	−.471	.496	10	.341	.674
11	.763	.761	10	.341	.606	11	.763	.818

Panel B. White Wines

2-Judge Composite: Judges 3 and 5 Accuracy of 2-Judge Composite: .525[a]			*2-Judge Composite: Judges 5 and 11 Accuracy of 2-Judge Composite: .665*[b]			*2-Judge Composite: Judges 2 and 4 Accuracy of 2-Judge Composite: .733*[c]		
3rd Judge Added	*Individual Accuracy of 3rd Judge*	*Accuracy of 3-Judge Composite*	*3rd Judge Added*	*Individual Accuracy of 3rd Judge*	*Accuracy of 3-Judge Composite*	*3rd Judge Added*	*Individual Accuracy of 3rd Judge*	*Accuracy of 3-Judge Composite*
1	.308	.461	1	.308	.586	1	.308	.549
2	.829	.683	2	.829	.736	3	.155	.528
4	.183	.499	3	.155	.605	5	.629	.741
7	.704	.606	4	.183	.660	7	.704	.775
8	−.211	.398	7	.704	.679	8	−.211	.434
9	.531	.566	8	−.211	.536	9	.531	.759
10	.462	.552	9	.531	.716	10	.462	.629
11	.707	.605	10	.462	.643	11	.707	.773

[a]50th percentile composite accuracy.
[b]75th percentile composite accuracy.
[c]90th percentile composite accuracy.

addition of Judge 8 to the 90th-percentile composite consisting of Judges 1 and 5: composite accuracy increases from .811 to .831 even though Judge 8's individual accuracy is only .195. The pairwise agreement results in Table 2 reveal that Judge 8 is uncorrelated with both Judges 1 and 5 (correlations of .107 and −.089, respectively), confirming Equation (2)'s implication about the importance of adding judges who are relatively uncorrelated with the existing composite.

IV. Conclusion

It is found that the accuracy of expert wine tasters' judgments can be improved substantially by forming composites of the individuals' judgments, and that most of the improvement that is possible can be achieved by including only two or three judges in the composite. While composite accuracy continues to improve with the inclusion of additional judges, the marginal benefit of including more than about three judges diminishes rapidly. These empirical findings are reinforced by analytical work that not only enables the impact of additional judges to be estimated but also suggests who they should be. The results have implications for better understanding the variability in the price-quality association across hedonic wine pricing studies, as well as for the practical issue of choosing judges to include in tasting panels that provide expert advice.

References

Armstrong, J.S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.

Ashenfelter, O. and Quandt, R. (1999). Analyzing a wine tasting statistically. *Chance*, 12, 16–20.

Ashton, R.H. (1986). Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes*, 38, 405–414.

Ashton, A.H. and Ashton, R.H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, 31, 1499–1508.

Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20, 451–468.

Benjamin, B.A. and Podolny, J.M. (1999). Status, quality, and social order in the California wine industry. *Administrative Science Quarterly*, 44, 563–589.

Broomell, S.B. and Budescu, D.V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, 74, 531–553.

Budescu, D.V. and Yu, H.T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, 20, 153–177.

Cicchetti, D.V. (2004a). Who won the 1976 blind tasting of French Bordeaux and U.S. cabernets? Parametrics to the rescue. *Journal of Wine Research*, 15, 211–220.

Cicchetti, D.V. (2004b). On designing experiments and analysing data to assess the reliability and accuracy of blind wine tastings. *Journal of Wine Research*, 15, 221–226.

Cicchetti, D.V. (2006). The Paris 1976 wine tastings revisited once more: comparing ratings of consistent and inconsistent tasters. *Journal of Wine Economics*, 1, 125–140.

Clemen, R.T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5, 559–583.

Clemen, R.T. and Winkler, R.L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research*, 33, 427–442.

Combris, P., Lecocq, S. and Visser, M. (1997). Estimation of a hedonic price equation for Bordeaux wine: Does quality matter? *The Economic Journal*, 107, 390–402.

Combris, P., Lecocq, S. and Visser, M. (2000). Estimation of a hedonic price equation for Burgundy wine. *Applied Economics*, 32, 961–967.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Dawes, R.M. (1970). An inequality concerning correlation of composites vs. composites of correlations. *Oregon Research Institute Methodological Note*, Vol. 1., No. 1.

Dawes, R.M. and Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin,* 81, 95–106.

Di Vittorio, A. and Ginsburgh, V. (1996). Pricing red wines of Medoc vintages from 1949 to 1989 at Christie's auctions. *Journal de la Société de Statistique de Paris*, 137, 19–49.

Einhorn, H.J. and Hogarth, R.M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance,* 13, 171–192.

Einhorn, H.J., Hogarth, R.M. and Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, 84, 158–172.

Ferrell, W.R. (1985). Combining individual judgments. In G. Wright (Ed.), *Behavioral Decision Making*. New York: Plenum Publishing Corporation.

Flores, B.E. and White, E.M. (1988). A framework for the combination of forecasts. *Journal of the Academy of Marketing Science*, 16, 95–103.

Gawel, R. and Gooden, P.W. (2008). Evaluation of the consistency of wine quality assessments from expert wine tasters. *Australian Journal of Grape and Wine Research*, 14, 1–8.

Ghiselli, E.E. (1964). *Theory of Psychological Measurement*. New York: McGraw-Hill.

Gibbs, M., Tapia, M. and Warzynski, F. (2009). Globalization, superstars, and reputation: theory & evidence from the wine industry. *Journal of Wine Economics*, 4, 46–61.

Gigone, D. and Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, 121, 149–167.

Goldwyn, C. and Lawless, H. (1991). How to taste wine (for fun and profit). *ASTM Standardization News*, 19, 32–37.

Hadj Ali, H., Lecocq, S. and Visser, M. (2008). The impact of gurus: Parker grades and *en primeur* wine prices. *The Economic Journal*, 118, 158–173.

Haeger, J.W. and Storchmann, K. (2006). Prices of American pinot noir wines: climate, craftsmanship, critics. *Agricultural Economics*, 35, 67–78.

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

Hogarth, R.M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, 21, 40–46.

Hogarth, R.M. and Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, 114, 733–758.

Hulkower, N. (2009). The judgment of Paris according to Borda, *Journal of Wine Research*, 20, 171–182.

Jones, G.V. and Storchmann, K. (2001). Wine market prices and investment under uncertainty: an econometric model for Bordeaux crus classés. *Agricultural Economics*, 26, 115–133.

Larrick, R.P. and Soll, J.B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, 52, 111–127.

Lawless, H., Liu, Y. and Goldwyn, C. (1997). Evaluation of wine quality using a small-panel hedonic scaling method. *Journal of Sensory Studies,* 12, 317–332.

Lecocq, S. and Visser, M. (2006). What determines wine prices: objective vs. sensory characteristics. *Journal of Wine Economics*, 1, 42–56.

Libby, R. and Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Performance*, 21, 121–129.

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

Makridakis, S. and Winkler, R.L. (1983). Averages of forecasts: Some empirical results. *Management Science*, 29, 987–996.
Oczkowski, E. (1994). A hedonic price function for Australian premium table wine. *Australian Journal of Agricultural Economics*, 38, 93–110.
Oczkowski, E. (2001). Hedonic wine price functions and measurement error. *The Economic Record*, 77, 374–382.
Quandt, R.E. (2006). Measurement and inference in wine tasting. *Journal of Wine Economics*, 1, 7–30.
Quandt, R.E. (2007). A note on a test for the sum of ranksums. *Journal of Wine Economics*, 2, 98–102.
Roberts, P.W. and Reagans, R. (2007). Critical exposure and price-quality relationships for new world wines in the U.S. market. *Journal of Wine Economics*, 2, 56–69.
Rudin, W. (1987). *Real and Complex Analysis*. New York: McGraw-Hill.
San Martin, G.J., Troncoso, J.L. and Brümmer, B. (2008). Determinants of Argentinean wine prices in the U.S. *Journal of Wine Economics*, 3, 72–84.
Schamel, G. and Anderson, K. (2003). Wine quality and varietal, regional and winery reputations: hedonic prices for Australia and New Zealand. *The Economic Record*, 79, 357–369.
Schnabel, H. and Storchmann, K. (2010). Prices as quality signals: evidence from the wine market. *Journal of Agricultural & Food Industrial Organization*, 8, 1–21.
Taber, G.M. (2006). *Judgment of Paris: California vs. France and the Historic 1976 Paris Tasting that Revolutionized Wine*. New York: Scribner.
Tellis, G.J. and Wernerfelt, B. (1987). Competitive price and quality under asymmetric information. *Marketing Science*, 6, 240–253.
Troncoso, J.L. and Aguirre, M. (2006). Price determinants of Chilean wines in the U. S. market: A hedonic approach. *Spanish Journal of Agricultural Research*, 4, 124–129.
Unwin, T. (1999). Hedonic price indexes and the qualities of wines. *Journal of Wine Research*, 10, 95–104.
Wallsten, T.S., Budescu, D.V., Erev, I. and Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making,* 10, 243–268.
Winkler, R.L. and Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society, Series A*, 146, 150–157.

Appendix

The effectiveness of composites of two or more judges for improving accuracy, shown in Tables 4 and 5 using a correlational accuracy measure (CORR), is revisited here with an alternative accuracy measure – the mean absolute percentage error (MAPE) between the "correct" ratings and those of the composites. The calculation of composite accuracy is identical to that described in the text for CORR, except that absolute errors, expressed as a percentage of Christian Millau's ratings, were calculated and the mean absolute percentage error was found for each composite. Note that while higher numbers indicate greater accuracy for CORR, lower numbers indicate greater accuracy for MAPE.

Prior research (e.g., Gigone and Hastie, 1997) suggests that correlational and distance-based accuracy measures are unlikely to produce different results in

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press

Table A1
Descriptive Statistics for MAPE: Red Wines

Number of Judges (k)	*Number of Composites of Size k*	*Mean*	*Variance*	*Low*	*Median*	*High*
1	–	32.34	0.0082	17.89	32.30	43.71
2	45	26.14	0.0053	16.38	24.94	42.57
3	120	23.52	0.0038	13.22	22.18	38.93
4	210	22.11	0.0028	12.45	21.29	35.93
5	252	21.25	0.0021	13.31	20.68	33.62
6	210	20.49	0.0015	13.03	19.84	31.74
7	120	20.01	0.0011	14.25	19.80	29.19
8	45	19.56	0.0007	14.72	19.20	25.83
9	10	19.17	0.0004	16.38	19.36	22.16
10	1	19.04	–	–	–	–

Table A2
Descriptive Statistics for MAPE: White Wines

Number of Judges (k)	*Number of Composites of Size k*	*Mean*	*Variance*	*Low*	*Median*	*High*
1	–	35.92	0.0117	24.67	32.65	54.08
2	45	30.16	0.0051	16.32	29.66	48.89
3	120	28.57	0.0032	17.97	28.10	43.64
4	210	27.79	0.0024	17.20	27.25	39.36
5	252	27.48	0.0017	19.00	27.35	36.72
6	210	27.36	0.0012	19.65	27.45	34.67
7	120	27.32	0.0008	20.33	27.47	32.78
8	45	27.32	0.0005	21.93	27.79	31.28
9	10	27.32	0.0002	24.35	27.55	29.73
10	1	27.32	–	–	–	–

judgment settings such as the present one, and indeed they do not. The results for MAPE, presented in Tables A1 and A2, are quite consistent with those presented earlier for CORR. For the reds, Table A1 reveals a mean MAPE of 32.34% for the 10 individual judges. In other words, the mean absolute distance between Millau's ratings and the ratings of the 10 judges is about 32 percent of Millau's ratings. Mean MAPE declines steadily from 32.34% when $k = 1$ to 19.04% when $k = 10$. The largest decline occurs from $k = 1$ to $k = 2$, after which mean MAPE declines at a decreasing rate – as with the mean CORR metric. Focusing on the maximum possible improvement as k increases from 1 to 10 (i.e., 13.30%, or 32.34% – 19.04%), the average 2-judge (3-judge) composite achieves 47% (66%) of this maximum. The "High" column of Table A1 (showing the *worst* composite at each level of k) depicts much greater improvement in accuracy as k increases. In fact, the *worst* composites of six or more judges produce lower MAPEs than the mean MAPE for a single judge. The "Low" column (showing the *best* composite at

each level of k) also generally declines as the size of the composite increases, and the variability of MAPE decreases as well.

The white wines (Table A2) differ from the reds in that the absolute errors are greater at all levels of k, whether we focus on the high, low or mean MAPE. However, the effectiveness of composites for decreasing the magnitude of the error is clear. As with the reds, the largest decline in mean MAPE occurs at $k = 2$. The average 2-judge (3-judge) composite achieves 67% (86%) of the maximum possible improvement as k increases from 1 to 10 (i.e., 8.60%, or 35.92% – 27.32%). The worst composites of $k \geq 6$ produce lower MAPEs than the mean single-judge MAPE, and even the *best* judgments can be improved by forming composites. Thus, in all respects the results based on MAPE are very similar to those based on CORR.

Finally, instead of focusing on "percentage of maximum possible improvement" as a way of expressing the power of composites, an alternative might be "percentage reduction of mean *individual* error." Armstrong (2001) used the latter measure in his analysis of 30 studies published between 1960 and 2000, almost all of which had relied on mean absolute error (MAE) or mean absolute percentage error (MAPE) as the measure of individual and composite accuracy. Across a wide range of judgment contexts, all involving experts, Armstrong found that mean individual MAE or MAPE was reduced by an average of 12.5% (range = 3–24%) across the studies by forming composites of various sizes. Examination of the "Mean" columns of Tables A1 and A2 reveals larger error reductions in the present context than those found by Armstrong. For the red wines, the reduction in mean individual error achieved by 2-judge (3-judge) composites is 19% (27%) and that achieved by the single 10-judge composite is 41%. The comparable error reductions for the whites are 16%, 20% and 24%, respectively.

https://doi.org/10.1017/S1931436100001577 Published online by Cambridge University Press