Journal of Wine Economics, Volume 7, Number 1, 2012, Pages 70–87
doi:10.1017/jwe.2012.6

Reliability and Consensus of Experienced Wine Judges: Expertise Within and Between?

Robert H. Ashton[a]

Abstract

This paper considers the levels of reliability and consensus of wine quality judgments found in studies of experienced wine judges. Both reliability, which concerns the similarity of repeat judgments of a particular wine by the *same* judge, and consensus, which concerns the similarity of judgments of a particular wine *across* judges, are necessary requirements for expertise in wine judging. Reliability and consensus levels found in wine judging are compared to those documented by a large body of research in six other fields: medicine, clinical psychology, business, auditing, personnel management, and meteorology. In all fields, including wine judging, reliability is greater than consensus. Both reliability and consensus are, on average, substantially lower in wine judging than in other fields, although tremendous variability exists across judges in every field. Overall, little support is found for the idea that experienced wine judges should be regarded as experts. (JEL Classification: C91)

I. Introduction

Recent research on wine judging raises questions about both the reliability and consensus of wine quality judgments made by experienced judges in blind tastings (e.g., Gawel and Godden, 2008; Hodgson, 2008, 2009a, 2009b). *Reliability,* an intraindividual notion, concerns the similarity of repeat judgments of the same wine by an individual judge, while *consensus,* an interindividual notion, concerns the similarity of the judgments of a particular wine between/among two or more independent judges.[1] Both reliability and consensus are necessary requirements for expertise in wine judging. Stated simply, the basic issues are the extent to which individual wine judges repeat their own judgments (which I label "expertise

[a] L. Palmer Fox Professor, Fuqua School of Business, Duke University.

[1] Terminology varies, especially in the wine literature. For example, "consistency" is often used to refer to intrajudge variability (Hodgson, 2008), and both "reliability" and "concordance" are often used to refer to interjudge variability (Cicchetti, 2004b; Hodgson, 2008). In the broader social science literature, in which the judgments of professionals in many fields have been investigated extensively, "reliability" refers to intrajudge variability, while "consensus" refers to interjudge variability. I use the latter terminology in this paper.

within") and the extent to which different wine judges agree in their judgments (which I label "expertise between").

Although the body of research on the reliability and consensus of experienced wine judges is small (and mostly recent), both the "within" and "between" aspects of judgment variability have been the subject of extensive research across many professional fields over many decades. In this paper, I review the few wine studies that exist and compare their results with those of a much larger sample of carefully controlled experimental studies that examine reliability and consensus in the fields of medicine, clinical psychology, business, auditing, personnel management, and meteorology. All the studies that I review quantify reliability and consensus using correlational measures. Reliability for each individual judge is measured as the correlation between repeat judgments of identical stimuli on two different occasions. Consensus is measured as the correlation between the judgments of identical stimuli by each *pair* of judges. Correlational measures are by far the most common way of quantifying reliability and consensus, and they offer the advantage of greater comparability of the levels of reliability and consensus across individuals, judgment tasks, and, ultimately, professional fields. I exploit this advantage by contrasting the level of reliability and consensus found in wine judging with that found in other fields.

Of course, professional judges—in any field—cannot be expected to achieve perfect reliability or perfect consensus, especially the latter, for many reasons. These include varying levels of attention to the task and motivation to perform well, differential ability and experience, focusing on different aspects of the phenomenon of interest, the absence of objectively "correct" or "best" answers in many settings, and the evolving and dynamic nature of the phenomenon being judged (Shanteau, 2001). As a result, it is difficult to make statements about the level of reliability and consensus that one should expect to find in a particular field. In controlled experimental settings, however, many factors that might naturally degrade reliability and consensus will likely operate to a lesser extent, so we might be seeing the various types of judges "at their best" in the research results. In any event, we can get a clear sense of the *relative* extent of reliability and consensus across fields, and that sense can help to inform our understanding of these critical aspects of wine judging.

Section II of the paper considers the roles that reliability and consensus play in the evaluation of professional judgment. It also describes the positive relationship that exists between reliability and accuracy and between consensus and accuracy, in settings where "correct answers" are available and thus accuracy can be measured. Section III consists of two parts. The first part summarizes the results of wine studies that examine intrajudge reliability and compares the results to those of 41 studies conducted in these other six fields. The second part summarizes the results of wine studies that examine interjudge consensus and compares the results to those of 46 studies conducted in these same fields. In addition, results from several studies that examine both reliability and consensus in the same study and with the same judges are briefly presented. Section IV presents a discussion and conclusion.

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

II. Accuracy, Reliability, and Consensus

The type of setting addressed here is that in which one or more individuals in a professional field, who might be "experts" in varying degrees, make professional judgments concerning specialized aspects of their field, and then communicate recommendations based on those judgments to people who use them as critical inputs in their decision making. Examples abound in such fields as medicine, business, and consumer decision making. Judgments and recommendations made in these settings can be highly consequential to both those who provide them (because of their effect on reputation-building) and those who receive them (because of their effect on decisions).

Those on the receiving end seek confidence in the recommendations they receive and, therefore, are interested in the quality of the professional judgments on which those recommendations are based. Ideally, the quality of professional judgments would be revealed by their *accuracy,* that is, their correspondence with an objectively measured external criterion that is independent of the professional and the judgments he or she makes. In many settings, however, an independent external criterion does not exist (or will not be known for a long time), and therefore judgment accuracy cannot be evaluated. In those settings, attention naturally turns to surrogate evaluation criteria such as intrajudge reliability and interjudge consensus, criteria that are necessary but not sufficient for establishing expertise (or at least for establishing that such judgments are "good enough" for practical purposes). Instead of being three separate features of professional judgment, however, accuracy, reliability, and consensus are closely related both theoretically and empirically, as explained below.

Researchers across many fields consider reliability a more fundamental requirement for expertise than consensus. Cicchetti (2004b) and Hodgson (2008, 2009b), for example, adopt this view in the field of wine judging: "What do we expect from expert wine judges? Above all, we expect [reliability], for if a judge cannot closely replicate a decision for an identical wine served under identical circumstances, of what value is his/her recommendation?" (Hodgson, 2009b, 241). A similar view prevails in medicine. In a setting involving judgments of disease severity, Einhorn (1974, 563) states, "With regard to intrajudge reliability, it should be obvious that unless the expert can reproduce his [judgments], there is little more that can be said in defense of his expertise." Similarly, in a setting involving the evaluation of coronary angiograms, Detre, Wright, Murphy and Takaro (1975, 985) state, "Although high intra- and interobserver agreement does not assure that the observer is right in his judgment, it is certain that he could hardly be right if he disagrees often with himself." Thus, intrajudge reliability is typically regarded as the most important requirement for expertise when the absence of objectively correct answers prevents a definitive determination of judgment accuracy.

It must be recognized, however, that reliability remains an important requirement for expertise even when objectively correct answers are available, and therefore judgment accuracy can be assessed, because of the positive relationship between reliability and accuracy. Theoretical work establishes that intrajudge reliability places an upper limit on the level of accuracy that can be achieved (e.g., Ghiselli, 1964; Lord and Novick, 1968). This fact is captured by Goldberg's (1970, 423) description of intrajudge reliability issues in the field of clinical psychology: "He 'has his days': Boredom, fatigue, illness, situational and interpersonal distractions all plague him, with the result that his repeated judgments of the exact same stimulus configuration are not identical. He is subject to all those human frailties which lower the reliability of his judgments below unity. And, if the judge's reliability is less than unity, there must be error in his judgments—error which can serve no other purpose than to attenuate his accuracy." Thus, intrajudge reliability is a necessary requirement for expertise both when the accuracy of professional judgment cannot be assessed and when it can.

It is worth noting that test-retest reliability is not the only type of intrajudge reliability that has been studied by judgment researchers. The other principal type, often called "linear consistency," concerns the extent to which a linear regression model estimated from the relationship between an individual's judgments and a set of underlying information items can reproduce the individual's judgments. This type of intrajudge reliability is one determinant of the ability of a linear regression model of the individual to produce accurate predictions of an external criterion. Linear-consistency and test-retest reliability are related (see Cooksey, 1996, 205–208) in that linear-consistency reliability is a function of test-retest reliability and the extent to which the individual's linear regression model captures the underlying judgment process, that is, the extent to which the individual's judgment process reflects the linearity and additivity assumptions that underlie regression (Lee and Yates, 1992). Because linear-consistency reliability confounds the effects of test-retest reliability with the effects of systematic departures from linearity and additivity, test-retest reliability is the more fundamental of the two.[2]

Although reliability is widely considered a more fundamental requirement of professional judgment than is consensus, as observed earlier, consensus is nevertheless extremely important. This is especially true in settings where correct answers do not exist (or will not be known within a reasonable period). Decisions must be made and actions must be taken even though the "correctness" of those decisions might never be known. Because agreement among the independent judgments of competent professionals is often an indispensible input to decisions and actions, consensus has emerged as an important criterion for evaluating judgment. As Hodgson (2008, 106) puts it in the wine context, "good judges agree with each other."

[2] A full treatment of linear-consistency reliability is beyond the scope of this paper. Although linear-consistency reliability has not been evaluated in the context of wine judging, scores of studies in other fields find a strong positive relation between this type of intrajudge reliability and accuracy—see, for example, the meta-analyses of Karelaia and Hogarth (2008) and Kaufmann and Athanasou (2009).

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

To the extent that interjudge agreement is considered a desirable feature of wine judging, it follows that ways of increasing such agreement are likely to be of interest. Indeed, Cicchetti (2004b, 221), in his discussion of research designs and data-analytic strategies for improving blind wine tastings, says "the goal is to reduce, as much as is possible, the extent of inter-judge variability in the evaluation of any given wine." Cicchetti goes further, however, making a bold suggestion that reducing interjudge variability should "[increase] the validity or accuracy of blind wine tasting" (221).

The idea that reducing interjudge variability (i.e., increasing consensus) will result in increased accuracy has been tested empirically by Ashton (1985) in two important business settings where correct answers exist. One setting involves sales predictions (a continuous judgment variable), in which Time, Inc., executives make quarterly predictions, over fourteen years, of the annual number of advertising pages that will be sold by *Time* magazine. The second setting involves predictions by independent auditors (CPAs) of whether a sample of business firms will or will not continue as "going concerns" (a dichotomous judgment variable) for the coming year. In both settings, a strong positive relationship is found between consensus and accuracy.[3] Ashton (1985, 185) concludes: "If an individual's predictions agree strongly with those of others in a group, then that individual will tend to be among the most accurate in the group. This conclusion also holds for *pairs* of individuals; that is, pairs who agree better also tend to be more accurate than other pairs. Similarly, individuals and pairs that exhibit low consensus tend to be less accurate than those exhibiting high consensus."

Ashton's (1985) finding of a strong positive relation between consensus and accuracy is bolstered by the results of Detre et al. (1975), who find a strong positive relationship between consensus and reliability. In a medical setting involving evaluations of coronary angiograms, these researchers document considerable variability in both reliability and consensus and, more important for present purposes, a clear relationship between the reliability of individual judges and how often they agree with other judges.

Despite results such as those of Ashton (1985) and Detre et al. (1975), consensus is sometimes viewed as a problematic criterion for evaluating judgment. Although it is difficult to dispute the notion that a professional judge should not "disagree with himself," it is often pointed out that even complete agreement among judges does not guarantee accuracy and that the lone dissenter among many judges

[3] This result held when consensus was measured in noncorrelational terms, e.g., as the mean absolute difference between the judgments of two individuals. The relationship between consensus and accuracy in the field of auditing is further explored by Davis, Kennedy and Maines (2000), Keasy and Watson (1989), and Pincus (1990). Kenny (1991) develops a more general theoretical model of consensus and accuracy in a broader context.

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

could, in fact, be correct. As Einhorn (1974, 570) states, "the history of science is replete with oddballs who did not agree with anyone, yet, were proved to be correct by subsequent events." Einhorn also observes, however, that the later discovery that the oddball was correct requires that a criterion other than consensus eventually become available, which will not be the case in many important judgment settings.

Perhaps a more troublesome aspect of consensus as a criterion for evaluating judgment is its potential dampening effect on learning: "Disagreements are often the route by which experts increase understanding of their field. By seeking out areas of disagreement between one another, experts explore the limits of their own knowledge and stretch their range of competency" (Weiss and Shanteau, 2004, 231). Thus, to the extent that agreement becomes the standard, the benefits of disagreement, alternative viewpoints, devils' advocates, and so on may be lost and learning may suffer. These potential drawbacks notwithstanding, the practical necessity for timely decisions and actions, as well as the positive relationship among consensus, reliability, and accuracy revealed by research, firmly establish consensus as an important criterion for evaluating judgment.

III. Results

A. *Judgment Reliability: Expertise Within?*

Correlational studies of the intrajudge reliability of experienced wine judges have been reported by Brien, May, and Mayo (1987), Gawel and Godden (2008), Gawel, Royal, and Leske (2002) and Lawless, Liu and Goldwyn (1997). Each study involves several judges who, in blind tastings, independently rate a number of wines and later re-rate those same wines. The researchers determine, separately for each judge, the correlation between the judge's first and second ratings. The results are summarized in Table 1, Panel A.

Brien et al. (1987) describe the results of four studies in which either 24 or 48 different wines were tasted—and re-tasted the same day or the following day—by either six or eight experienced judges. Intrajudge reliability varies greatly, ranging from .16 to 1.00. On average, reliability is fairly high, with mean reliability across the four studies ranging from .45 to .74. Note that the mean reliabilities in Studies 2 and 5, in which the repeat tastings occurred the same day (.73 and .74), are considerably greater than those in Studies 3 and 4, in which the repeat tastings occurred one day later (.45 and .54).

Lawless et al. (1997) report a study in which four panels of judges tasted—and re-tasted less than an hour later—14 different wines. Three of the panels were experienced wine tasters (Panels CB, G, and PB), while the fourth panel was described as wine consumers (Panel C). The range of intrajudge reliability across the individual judges is $-.03$ to .85, while the range of mean reliability across the

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

Table 1
Summary of Studies Investigating Judgment Reliability

Panel A. Wine Studies

	Reliability		
Study	*Mean*	*Lowest*	*Highest*
Brien, May and Mayo (1987)			
Study 2	.73	.39	.91
Study 3	.45	.35	.59
Study 4	.54	.16	.76
Study 5	.74	.56	1.00
Lawless, Liu, and Goldwyn (1997)			
Panel CB	.53	.10	.76
Panel G	.61	.32	.85
Panel PB	.42	.10	.80
Panel C	.31	−.03	.69
Gawel, Royal, and Leske (2002)			
Published data	.46	−.49	.98
Unpublished data	.40	–	–
Gawel and Godden (2008)			
Reds	.45	−.39	.97
Whites	.35	−.42	.97
Mean reliability across studies	.50		

Panel B. Studies in Other Fields

Field	*Number of studies*	*Mean reliability*
Meteorology	3	.91
Business	3	.83
Auditing	10	.82
Personnel management	13	.76
Medicine	6	.76
Clinical psychology	6	.70

individuals in the four panels is .31 to .61.[4] The consumer panel produces lower reliabilities than the three experienced panels (a mean of .31 for the former and means of .53, .61, and .42 for the latter).

A particularly interesting aspect of the Lawless et al. (1997) results is that the reliability of the *mean ratings* of the individuals in each panel is much higher than the mean reliability of the panel's individual members. To illustrate, consider Panel CB, which has six members. The mean reliability reported for Panel CB in Table 1 (.53) is the mean of the six judges' individual reliability values, consistent with the notion that reliability is an intraindividual phenomenon. In addition to quantifying these six individual *reliability values,* Lawless et al. also calculate the mean of the six

[4] The numbers reported here and in Table 1 are estimated from figure 1 in Lawless et al. (1997).

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

judges' *ratings* of each wine, on both the initial and repeat tastings, and then determine the correlation between these mean ratings. The resulting correlation (.90) is much higher than the mean of the six judges' individual reliability values (.53).[5] The superiority of mean, or composite, judgments vis-à-vis those of the average individual in the composite has been demonstrated in many settings, including wine judging (Ashton, 2011).

Gawel et al. (2002) report a study in which 42 experienced judges tasted a wine that had been aged in four different types of oak. Instead of rating overall quality, however, the judges rated the intensity of eight different characteristics of the wine (e.g., spice, butter, and texture). Only average reliabilities across the eight characteristics are reported. Again, there is tremendous variability across judges, with a mean reliability of .46. Gawel et al. (2002) also refer to unpublished data from 225 experienced tasters that reveal a mean intrajudge reliability of .40, although they provide no further information.

Gawel and Godden (2008) report results from tastings involving 571 experienced judges who tasted an average of 23 reds and 23 whites, with duplicates tasted two or three days later. Again, great variability across judges is evident, with mean intrajudge reliability of .45 for the reds and .35 for the whites. When the reliability of three-judge panels was evaluated, it was found to be substantially greater than the mean reliability of the individual judges—consistent with the earlier results of Lawless et al. (1997).

Finally, Hodgson (2008) reports some fascinating results from the California State Fair Wine Competitions of 2005 to 2008. Hodgson's results concern four triplicate samples that were judged by 16 panels of four judges each. Both of the repeat samples were tasted in the same tasting flight and were poured from the same bottle as the original sample. As Hodgson (2008, 106) explains: "The overriding principle was to design the experiment to maximize the probability in favor of the judges' ability to replicate their scores." Unlike in earlier studies, a correlational measure was not used in this study to quantify judge reliability; instead, the judges awarded medals to each wine (Gold, Silver, Bronze, or No Award), and the reported results concern the judges' ability to replicate their own awards. The key finding is that the judges awarded the same medal only about 18 percent of the time—and this usually occurred for wines that received No Award. Moreover, in many instances a judge awarded Gold to one of the triplicates and Bronze (or No Award) to another.

Mean reliability across all the wine studies in Table 1 is .50. How does this compare to judgment reliability in other fields? In an earlier paper, I analyzed

[5] The same result obtains for the other three panels of judges: the correlations between the mean *ratings* of the panel (.89, .82, and .52 for Panels G, PB, and C, respectively) are much higher than the mean correlations produced by the individuals who are included in the panel (.61, .42, and .31).

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

published research on the reliability of professional judgment in the fields of meteorology, medicine, clinical psychology, personnel management, business, and auditing (Ashton, 2000). Fifty studies across these six fields were identified, 41 of which measured reliability as the correlation between repeat judgments of identical stimuli by each judge. All 41 correlational studies focus on professional judges who make a series of judgments in the domain of their everyday experience (as opposed to, say, college students responding to abstract and unfamiliar tasks to fulfill a course requirement).

The meteorological studies concerned forecasts of atmospheric events such as microbursts and hail. The medical studies involved professionals such as pathologists and radiologists evaluating the severity of conditions such as gastric ulcers and Hodgkin's disease. The clinical psychology studies concerned the evaluation of traits such as intelligence and sociability. The personnel management studies concerned the evaluation of various dimensions of work-related behaviors, typically for selection or promotion purposes. The business studies concerned financial analysis and taxation. Several studies involved the professional field of auditing. Because the nature of professional judgment in auditing may be unfamiliar to readers of this journal, the Appendix provides a brief explanation of the critical importance of judgment in auditing.

Judgment reliability varied substantially across individual judges in these studies. The mean reliability that emerged in each of the six fields is reported in Table 1, Panel B. Mean reliability ranges from .91 in meteorology to .70 in clinical psychology—vis-à-vis a mean of .50 for the wine studies. (I defer until Section IV a consideration of why reliability in wine judging might reasonably be expected to be lower than in other fields.)

My earlier analysis (Ashton, 2000) identified three features of the overall body of results that may provide useful perspective in the wine context. First, reliability decreased with greater time between the original judgment and the repeat judgment, which is also seen in the Brien et al. (1987) study of wine judging. Second, group discussion among two or more individual judges had the effect of increasing reliability; a similar effect is seen in the superior reliability of the judge *panels* in Gawel and Godden (2008) and Lawless et al. (1997). Finally, reliability was inversely related to the difficulty of the judgment task; this, too, has its counterpart in studies of wine judging—for example, the clear tendency for reliability to be greater for wines at each end of the quality scale than for those in the middle (e.g., Hodgson, 2008).

B. *Judgment Consensus: Expertise Between?*

Correlational studies of the interjudge consensus of experienced wine judges have been reported by Ashton (2011), Baker and Amerine (1953), Brien et al. (1987), Cicchetti (2006a, 2006b), and Hodgson (2009a). Each study involves several judges who, in blind tastings, independently rate a number of wines. The researchers

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

Table 2
Summary of Studies Investigating Judgment Consensus

Panel A. Wine Studies

	Consensus		
Study	*Mean*	*Lowest*	*Highest*
Baker and Amerine (1953)			
Reds	.39	.07	.90
Whites	.58	.44	.75
Brien, May, and Mayo (1987)			
Study 2—Occasion 1	.45	−.09	.79
Study 2—Occasion 2	.37	−.40	.84
Cicchetti (2004a, 2006b)			
Reds	.22	.05	.93
Whites	.36	.09	.99
Ashton (2011)			
Reds	.16	−.62	.97
Whites	.44	−.07	.94
Hodgson (2009a)	.11	−.02	.33
Mean consensus across studies	.34		

Panel B. Studies in Other Fields

Field	*Number of studies*	*Mean consensus*
Meteorology	4	.75
Personnel management	6	.65
Auditing	23	.61
Medicine	3	.56
Business	8	.49
Clinical psychology	2	.37

determine, for each pair of judges, the correlation between their ratings. The results are summarized in Table 2, Panel A.

Baker and Amerine (1953; cited in Brien et al., 1987) report results from five experienced judges who evaluated 13 reds and 17 whites over multiple sessions, with four or five wines per session. The results reveal greater mean consensus for the whites (.58) than for the reds (.39). Substantial variability in consensus exists across pairs of judges, ranging from .44 to .75 for the whites and from .07 to .90 for the reds.

One of Brien et al.'s (1987) four reliability studies (described above) also examined consensus. Interjudge correlations are reported for both the first occasion on which the wines were tasted and the second occasion (later the same day). Mean consensus is lower in the repeat tasting than in the first (.37 vs. 45), and the range of consensus across judges is wider (−.40 to .84 vs. −.09 to .79).

Other evidence on the consensus of wine judgments comes from two analyses of the famous 1976 Paris tasting of California and French wines that revolutionized the

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

wine world.[6] Eleven experts (nine of them French) tasted ten reds (six California and four French) and ten whites (again, six California and four French). Although much has been written about who "won" the tasting (e.g., Ashenfelter and Quandt, 1999; Cicchetti, 2004a; Hulkower, 2009; Lindley, 2006; Quandt, 2006, 2007), my concern here is the extent to which the 11 judges agreed in their judgments of the wines. Cicchetti (2004a, 2006b), using the intraclass correlation coefficient as the measure of judge consensus, finds an overall consensus level of .22 for the reds and .36 for the whites. Ashton (2011), using the Pearson correlation as the measure of consensus, reports similar results: mean consensus of .16 for the reds and .44 for the whites. Both analyses report substantial variability in consensus across pairs of judges.

Hodgson (2009a) analyzed 4,167 wines that were entered in 13 major U.S. wine competitions in 2003. Several of his results speak to the degree of consensus in wine quality judgments across the competitions. First, 106 of the 375 wines that were entered in five competitions received Gold medals in one competition, but only 20 of these 106 received a second Gold medal and only six of these 20 received a third. None of the 375 received Gold medals in more than three competitions. Second, only 132 of the 3,347 wines that were entered in two or more competitions received the *same* medal in all competitions entered (and this almost always occurred in just two competitions). Finally, of the 2,440 wines that were entered in more than three competitions, 1,142 received at least one Gold; however, 957 of these 1,142 failed to receive *any* medal in at least one competition.

Hodgson (2009a) developed a correlational measure of consensus by first assigning numerical scores to the various medals and then computing correlations between the scores received by wines in each pair of competitions. With 13 competitions, there are 78 such pairwise measures. The mean correlation is .11, with a range of – .02 to .33. This clearly reflects poor consensus across the competitions, and most of the consensus that existed concerned wines awarded Bronze medals or No Awards. Hodgson (2009a, 5) concluded that "wine judges concur in what they do not like but are uncertain about what they do," consistent with his earlier findings (Hodgson, 2008) concerning intrajudge reliability.[7]

Lawless et al. (1997), whose intrajudge reliability results are reported in Table 1, also examined consensus. They did so, however, by focusing on the *mean* judgments of each of the four panels, not on the judgments of the individual members. As noted earlier, mean judgments result in inflated reliability values—and the same is true for consensus values. Lawless et al. found that the three experienced panels agreed much more with one another (correlations of .66, .75, and .77) than with the consumer panel (correlations of .33, .44, and .46).

[6] Taber (2006) provides a fascinating and informative account of the 1976 event.

[7] Cliff and King (1997), using a different methodology, report similar results: Judges agree much more on wines perceived to be at both the low and high end of the quality scale than on wines perceived to be of moderate quality.

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

Finally, Quandt (2006) summarizes some consensus results from 92 tastings conducted by the eight members of the Liquid Assets Wine Group. Instead of pairwise correlations among tasters, however, Quandt reports Kendall's coefficient of concordance (W), a measure of the overall concordance among the judges' ratings. Kendall's W is statistically significant at the .05 (.10) level for 49 percent (57 percent) of the tastings, indicating that "substantial agreement existed among judges more than half the time" (Quandt, 2006, 16).

Mean consensus across all the wine studies in Table 2 is .34, substantially below mean reliability across wine studies of .50. As is the case with reliability, it is of interest to compare the level of consensus found in wine judging to that found in other fields. To my knowledge, there is no comprehensive review of consensus studies comparable to Ashton's (2000) review of reliability studies. However, my recent search of the literature identified 46 studies across the same six professional fields included in Ashton (2000) that report consensus results using a correlational measure.[8] The types of judgments examined in each field are the same as those described above for the reliability studies, with the exception of studies in business; the eight consensus studies in business settings examine a wider range of issues than do the three reliability studies (including sales predictions, actuarial judgments, and predictions of stock prices). As in the reliability studies in Ashton (2000), all the consensus studies focus on professional judges who make a series of judgments in the domain of their everyday experience.

Table 2, Panel B, reports the mean consensus that emerged in each of the six fields. Mean consensus ranges from .75 in meteorology to .37 in clinical psychology —vis-à-vis a mean of .34 for the wine studies. (I defer until Section IV a consideration of why consensus in wine judging might reasonably be *expected* to be lower than in other fields.) Comparing mean reliability (Table 1) and mean consensus (Table 2) within fields reveals that consensus is lower than reliability in all fields, often substantially so, indicating that judges in all fields agree more with themselves than with others.

It should be recognized, however, that the mean within-field reliability and consensus results reported in Tables 1 and 2 are not completely comparable because the set of reliability studies in Table 1 differs somewhat from the set of consensus studies in Table 2 (i.e., some studies examine only reliability, some examine only consensus, and some examine both). Fortunately, many of these studies evaluate both reliability and consensus in the same study and with the same judges, allowing a direct comparison of reliability and consensus. The results, shown in Table 3, confirm that mean reliability is substantially greater than mean consensus in all fields. The *difference* between mean reliability and mean consensus ranges from .12 (.89 vs. .77) in meteorology to .36 (.73 vs. .37) in clinical psychology. (This compares

[8] Various subsets of these studies are described by Bédard and Chi (1993), Bouwman and Bradley (1997), Broomell and Budescu (2009), Shanteau (2001), and Wright (1988).

Table 3
Summary of Studies Investigating Both Reliability and Consensus in the Same Study

Field	*Number of studies*	*Mean reliability*	*Mean consensus*
Meteorology	3	.89	.77
Personnel management	6	.83	.65
Auditing	9	.81	.67
Business	3	.78	.58
Wine	1	.73	.41
Clinical psychology	2	.73	.37
Medicine	2	.70	.45

to a difference of .32 (.73 vs. .41) in the single wine study that examines both reliability and consensus.) Finally, examination of the mean within-study levels of reliability and consensus in the 25 non-wine studies reveals no case in which mean consensus exceeds mean reliability.

IV. Discussion and Conclusion

Both intrajudge reliability and interjudge consensus of experienced wine judges are found to be substantially below reliability and consensus in other fields. Quantified in correlational terms, mean reliability across published wine studies is .50 while mean consensus is .34. Moreover, reliability and consensus vary widely across studies (and across individual judges in a single study), with some judges performing well and others performing poorly. Two questions immediately arise: (1) Why are the *mean* levels of reliability and consensus so much lower among experienced wine judges than among judges in other fields? (2) What accounts for the great *variability* in reliability and consensus across wine judges?

On the first question, it is easy to imagine valid reasons that reliability and consensus in wine judging would be lower than in many other fields. At the risk of stating the obvious, foremost among them is that wine judging is inherently more subjective. Whereas professional judgment in meteorology, medicine, or business, for example, is based largely on relatively objective inputs (such as barometric pressure, x-ray results, and economic data), wine judging involves the senses of sight, smell, and taste. Thus, wine judging is not simply a matter of passively receiving some objective facts about bouquet, clarity, finish, and so forth and then weighting and combining these facts into an overall judgment of quality (which itself possesses a sizable subjective component).

The second question—concerning variability in reliability and consensus across wine judges—is more difficult. In general, however, a useful way of understanding the sources of differential performance across judges—in any field—is to focus on features of the judge, features of the judgment task, and the "interaction" between features of the judge and features of the task (Fischhoff, 1982). By interaction, I mean the extent to which there is a "match" (or a "mismatch") between judge and task.

Considering features of the judge (and assuming that the judge is *motivated* to perform well), voluminous research establishes that differences in ability, experience, and knowledge result in differential judgment performance (e.g., Ashton, 1999; Einhorn and Hogarth, 1981; Schmidt and Hunter, 1992). In the wine context, differences in *preferences* (which may result in part from differences in experience and knowledge and in part from past emotional associations involving particular wines) must also be considered, as must *biological* characteristics, such as differential sensitivity to smells and tastes (e.g., Bartoshuk, 1993; Goode, 2008). Considering features of the judgment task, a multitude of factors are involved in the task of blind wine tasting that might influence the overall results, including those with respect to intrajudge reliability and interjudge consensus (e.g., Amerine and Roessler, 1983; Goldwyn and Lawless, 1991). Examples include the types of wines tasted (and the range of types, if more than one), the number of wines of each type, the order in which they are tasted, the number of tasting flights, and the time between flights.

Such features of the judge and the judgment task surely account for much of the variability in reliability and consensus found across experienced wine judges. However, acquiring a deep understanding of differential performance across wine judges is likely to be more complex than identifying isolated features of judges and judgment tasks that are relevant. The extent to which relevant features of the judge are consonant with relevant features of the task (i.e., the extent to which there is a "match" between judge and task) is likely to be important as well.

To illustrate, imagine a blind tasting of red Bordeaux and red Burgundy involving four judges. Judges 1 and 2 have an affinity for Bordeaux, but not for Burgundy. Such affinity could be the result of greater experience or knowledge, stronger emotional associations, or heightened sensitivity and discriminability with respect to the smells and tastes of Bordeaux. In contrast, Judges 3 and 4 have the opposite affinity—for Burgundy, but not for Bordeaux. I conjecture that Judges 1 and 2 will exhibit greater intrajudge reliability when they taste Bordeaux (a match between judge and task) than when they taste Burgundy (a mismatch between judge and task) and that Judges 3 and 4 will exhibit the opposite pattern of results. Similarly, I conjecture that the Judge 1/Judge 2 pair and the Judge 3/Judge 4 pair will exhibit greater interjudge consensus than will the remaining four pairwise combinations of judges. They key point in this stylized example is that differential levels of reliability and consensus will not be determined solely by features of either the judge or the task in isolation but also by the extent to which those features match one another.

The empirical validity (and practical usefulness) of the various judge and task features mentioned above, as well as the notion that the "match" between judge and task is important in understanding the performance of experienced wine judges, can

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

only be settled by research. Existing studies on the reliability and consensus of experienced wine judges were not designed or conducted in a way that allows the sources of differential performance to be understood. I hope the results reviewed in this paper will provide a benchmark for future studies that take a systematic approach to understanding why reliability and consensus in wine judging are lower than in other fields *and* the sources of differential performance across experienced judges.

References

Amerine, M.A., and Roessler, E.B. (1983). *Wines: Their sensory evaluation*. New York: W.H. Freeman.

Arens, A.A., Elder, R.J., and Beasley, M.S. (2005). *Auditing and assurance services*. Upper Saddle River, NJ: Prentice Hall.

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

Ashton, A.H. (1985). Does consensus imply accuracy in accounting studies of decision making? *Accounting Review*, 60, 173–185.

Ashton, R.H. (1999). Enriching the "expertise paradigm" of accounting research: Conscientiousness, general cognitive ability, and goal orientation. *Advances in Accounting Behavioral Research*, 2, 3–14.

Ashton, R.H. (2000). A review and analysis of research on the test-retest reliability of professional judgment. *Journal of Behavioral Decision Making*, 13, 277–294.

Ashton, R.H. (2011). Improving experts' wine quality judgments: Two heads are better than one. *Journal of Wine Economics*, 6, 160–178.

Ashton, R.H., and Ashton, A.H. (1995). Perspectives on judgment and decision-making research in accounting and auditing. In R.H. Ashton and A.H. Ashton (Eds.), *Judgment and decision-making research in accounting and auditing*. New York: Cambridge University Press. Pages 3–5.

Baker, G.A., and Amerine, M.A. (1953). Organoleptic ratings of wines estimated from analytical data. *Food Research*, 18, 381–389.

Bartoshuk, L.M. (1993). The biological basis of food perception and acceptance. *Food Quality and Preference*, 4, 21–32.

Bédard, J., and Chi, M.T.H. (1993). Expertise in auditing. *Auditing: A Journal of Practice & Theory*, 12(Supplement), 21–45.

Bouwman, M.J., and Bradley, W.E. (1997). Judgment and decision making, part II: Expertise, consensus and accuracy. In V. Arnold and S.G. Sutton (Eds.), *Behavioral accounting research: Foundations and frontiers*. Sarasota, FL: American Accounting Association. Pages 89–133.

Brien, C.J., May, P., and Mayo, O. (1987). Analysis of judge performance in wine-quality evaluations. *Journal of Food Science*, 52, 1273–1279.

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

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

Cicchetti, D.V. (2006b). The 1976 blind wine tastings: On the consistency of tasters from chardonnays to cabernets. Vineyard Data Quantification Society (www.vdqs.net).

Cliff, M.A., and King, M.C. (1997). The evaluation of judges at wine competitions: The application of eggshell plots. *Journal of Wine Research*, 8, 75–80.

Cooksey, R.W. (1996). *Judgment analysis*. San Diego: Academic Press.

Davis, E.B., Kennedy, S.J., and Maines, L.A. (2000). The relation between consensus and accuracy in low-to-moderate accuracy tasks: An auditing example. *Auditing: A Journal of Practice & Theory*, 19, 101–121.

Detre, K.M., Wright, E., Murphy, M.L., and Takaro, T. (1975). Observer agreement in evaluating coronary angiograms. *Circulation*, 52, 979–986.

Einhorn, H.J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology*, 59, 562–571.

Einhorn, H.J., and Hogarth, R.M. (1981). Rationality and the sanctity of competence. *Behavioral and Brain Sciences*, 4, 334–335.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press. Pages 422–444.

Gawel, R., and Godden, 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.

Gawel, R., Royal, A., and Leske, P. (2002). The effect of different oak types on the sensory properties of Chardonnay. *Australian and New Zealand Wine Industry Journal*, 17, 14–20.

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

Goldberg, L.R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, 73, 422–432.

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

Goode, J. (2008). Experiencing wine: Why critics mess up (some of the time). In F. Allhoff (Ed.), *Wine & philosophy: A symposium on thinking and drinking*. Malden, MA: Blackwell. Pages 137–153.

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

Hodgson, R.T. (2009a). An analysis of the concordance among 13 U.S. wine competitions. *Journal of Wine Economics*, 4, 1–9.

Hodgson, R.T. (2009b). How expert are "expert" wine judges? *Journal of Wine Economics*, 4, 233–241.

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

Karelaia, N., and Hogarth, R.M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, 134, 404–426.

Kaufmann, E., and Athanasou, J.A. (2009). A meta-analysis of judgment achievement as defined by the lens model equation. *Swiss Journal of Psychology*, 68, 99–112.

Keasey, K., and Watson, R. (1989). Consensus and accuracy in accounting studies of decision making: A note on a new measure of consensus. *Accounting, Organizations and Society*, 14, 337–345.

Kenny, D.A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, 98, 155–163.

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.

Lee, J.W., and Yates, J.F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin*, 12, 363–377.

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

Lord, F., and Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Pincus, K.V. (1990). Audit judgment consensus: A model for dichotomous decisions. *Auditing: A Journal of Practice & Theory*, 9, 1–20.

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.

Schmidt, F.L., and Hunter, J.E. (1992). Development of a causal model of processes determining job performance. *Current Directions in Psychological Science*, 1, 89–92.

Shanteau, J. (2001). What does it mean when experts disagree? In E. Salas and G. Klein (Eds.), *Linking expertise and naturalistic decision making*. Hillsdale, NJ: Erlbaum. Pages 229–244.

Taber, G.M. (2006). *Judgment of Paris: California vs. France and the historic 1976 Paris tasting that revolutionized wine*. New York: Scribner.

Weiss, D.J., and Shanteau, J. (2004). The vice of consensus and the virtue of consistency. In K. Smith, J. Shanteau, and P. Johnson (Eds.), *Psychological investigations of competence in decision making*. Cambridge: Cambridge University Press. Pages 226–240.

Wright, W.F. (1988). Audit judgment consensus and experience. In K.R. Ferris (Ed.), *Behavioral accounting research: A critical analysis*. Columbus, OH: Century VII. Pages 305–328.

Appendix

Several of the reliability and consensus studies in Tables 1–3 concern judgments made in the field of auditing, a field that might be unfamiliar to readers of this journal. This Appendix provides some perspective on the critical role that professional judgment plays in auditing.

Briefly stated, auditing provides independent assurance concerning important disclosures provided by business organizations whose ownership shares are publicly held. Such organizations are required to disclose to current and potential investors and creditors substantial information about their past financial performance and current financial condition. Because this information is generated and disclosed by managers of the organization itself, who have strong incentives to portray the results favorably, and because external parties have limited access to such information via other channels, regulatory bodies in both the public and private sectors require the information to be examined by a firm of auditors, or certified public accountants (CPAs), who are independent of the reporting organization.

Auditors examine the reporting organization's financial disclosures and underlying systems and records to judge whether the disclosures are fairly presented in

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press

accordance with measurement and disclosure standards adopted by government agencies (e.g., the U.S. Securities and Exchange Commission) and the financial community more generally. Auditors collect and evaluate information that bears on this issue, and they use it as input to several component judgments that, when aggregated, suggest whether the organization's claim of fair presentation is likely to be tenable.

Audit judgments fall into two broad categories—investigation and reporting. *Investigation* judgments concern (1) the likelihood that errors or irregularities have occurred in the organization's processing of financial information and that the organization's own controls would have prevented or detected them, (2) the extent to which errors or irregularities that may have occurred and not been detected are important enough to require close scrutiny by the auditor, and (3) the extent to which evidence collection should be expanded in response to ongoing findings from the audit. *Reporting* judgments concern how best to fulfill the auditors' obligation to report (to the public) the results of their investigation. Auditors' most important reporting options are the standard and modified reports. A standard report provides assurance that the organization's financial disclosures are indeed fairly presented in accordance with accepted measurement and disclosure standards. A modified report, in contrast, signals that the organization's claim of fair presentation is unlikely to be tenable, and it provides an explanation of the circumstances or events that call fair presentation into question.

Auditors' investigation and reporting judgments are made in a setting that imposes significant costs on the various parties from legal, economic, and regulatory sources. Investors, creditors, suppliers, employees, and others can be harmed if auditors fail to detect errors or irregularities or fail to provide adequate disclosure of an organization's financial problems. The organization itself can be harmed if auditors mistakenly believe that they have found errors or irregularities or report that the organization has not provided adequate disclosure when, in fact, it has. Thus, the field of auditing is ripe for the study of professional expertise.[9]

[9] Arens, Elder and Beasley (2005), Ashton and Ashton (1995), and Bédard and Chi (1993) provide detailed accounts of the professional judgment issues faced by auditors.

https://doi.org/10.1017/jwe.2012.6 Published online by Cambridge University Press