beverages

MDPI

Article

Guidelines for Assessing Enological and Statistical Significance of Wine Tasters' Binary Judgments

Dom Cicchetti

Department of Biometry, Yale University School of Medicine, Box 317, North Branford, CT 06471, USA; dom.cicchetti@yale.edu; Tel.: +1-203-488-6563

Academic Editor: Edgar Chambers IV
Received: 28 September 2017; Accepted: 31 October 2017; Published: 3 November 2017

Abstract: The purpose of this article is to assess the reliability and accuracy (validity) of hypothetical binary tasting judgments in an oenological framework. The model that is utilized allows for the control of a wide array of variables that would be exceedingly difficult to fully control in the typical oenological investigation. It is shown that results that are judged to be oenologically significant are uniformly judged to be statistically significant as well, whether the level of Wine Taster agreement is set at 70% (Fair); 80% (Good), or 90% (Excellent). However, in a number of instances, results that were statistically significant were not enologically significant by standards that are widely accepted and utilized. This finding is consistent with the bio-statistical fact that given a sufficiently large sample size, even the most trivial of results will prove to be statistically significant. Consistent with expectations, multiple patterns of 80% (Good) and 90% (Excellent) agreement tended to be both statistically and enologically significant.

Keywords: hypothetics; enothetics; reliability; validity; accuracy

1. Introduction

The objective of this research report is to present a detailed analysis of the relationship between enological and statistical significance of research results as they both relate to the reliability and accuracy of Wine Tasters' hypothetical judgments. Reliability is defined here as the extent to which any given binary wine judgment is interchangeable with that of another wine judge (e.g., agreement that a wine is of excellent quality). The greater the extent to which this occurs, the higher the level of reliability.

The accuracy or validity of a hypothetical binary decision refers to the extent to which any pair of Wine Tasters renders the same correct judgment, for example, they both agree, correctly, that the wine is oaked or unoaked, or that the grape varietal is Syrah rather than Grenache. With respect to enological research investigations and scientific investigations more broadly, it is a well-known fact that uncontrolled variables can serve to compromise or call into question the accuracy of the reported findings.

In a previous study, a method was introduced, in an enological context, to address this vexing, albeit critical issue. Referred to as hypothetics, or enothetics in the current research context, the method allows investigators to begin to answer what findings would occur if it were indeed possible to control for variables that are often very difficult or, in some situations, impossible to control in the typical research study. A distinct advantage of the method is that it can also serve to highlight findings that would have become apparent if it were possible to control relevant variables. For example, in a recent enological investigation, it was shown that overall accuracy is a very poor measure of binary wine judgments, such as whether a wine is oaked or not [1]. Specific measures of judgmental accuracy, such as Sensitivity (Se), Specificity (Sp), Predicted Positive Accuracy (PPA) and Predicted Negative

Accuracy (PNA) were found to be much more useful measures of wine judgments than overall accuracy. The bio-statistical importance of such findings has relevance in designing future enological research investigations and in the design of scientific studies more generally.

2. The Role of Chance in Scientific Research

With respect to both reliability and accuracy of judgment, it should be noted that in any given inter-taster experiment, whether blind or open, a certain amount of measureable agreement will occur on the basis of chance alone. Therefore, appropriate reliability statistics all present as chance-corrected coefficients. This holds true quite irrespective of whether the statistics were designed for nominal variables, such as binary wine judgments [2]; or the Sensitivity-Specificity model—e.g., in an enological context [1]; ordinal variables [3]; or variables that are measured on interval or ratio scales [4–6].

For binary variables, the level of agreement expected on the basis of chance alone is calculated in the exact same manner as for the venerable and most familiar chi-square(d) statistic; and as applied correctly by Cohen [2] in the development of his kappa statistic, which was recently empirically verified [7].

3. Criteria for Assessing Levels of Practical Significance of the Reliability of Wine Judgments

There are currently three sets of published guidelines that were developed specifically for assessing the degree of the clinical or practical significance of a binary diagnostic judgment, as opposed to its level of statistical significance. In wine research, it would seem useful to refer to the term as enological significance. Three sets of criteria have been published [8–12]. As one might expect, the term clinical significance has its roots in bio-behavioral research, notably in nosology or diagnostic specialty areas. Practical significance is also synonymous with the phrase strength of agreement [8] and also with the concept of Effect Size (ES), as introduced by Cohen [13].

4. The Landis and Koch (1977), Fleiss (1981) and Cicchetti (1994) Enological Criteria

The Landis and Koch guidelines [8] contain six ordinal categories of increasing gradations of Strength of Agreement. These guidelines would seem particularly useful in an enological context in which wine experts were teaching less well experienced wine tasters to appreciate some of the nuances of wine judgments and then testing their reliability levels with the wine experts, at specific time points in the training exercise.

The Fleiss guidelines [10] consist of three ordinal categories of clinical significance; they would be applicable if the primary emphasis was to tri-chotomize wine judgments into unacceptable (Poor); acceptable (Fair or Good) and highly acceptable (Excellent).

Finally, the Cicchetti guidelines [12] consist of four ordinal categories of clinical significance. It would be most applicable if one were to relate them to clinical diagnoses, as in a nosological investigation of Autism [14] or in the present enological research context. In comparing the Fleiss et al. guidelines to those of Cicchetti and Sparrow, the latter make a distinction between Fair and Good, thereby forming four categories rather than three.

It should also be noted that because of the demonstrated equivalence between k, k_w and the intra-class correlation coefficient (*ICC*), the criteria apply regardless of the type of variable under investigation. First, Fleiss [15] demonstrated the mathematical equivalence between Cohen's kappa statistic (k) for nominal binary variables and the intra-class correlation coefficient for variables deriving from interval scales; and, secondly, Fleiss and Cohen [16] demonstrated the mathematical equivalence between Cohen's weighted kappa coefficient [3] and the *ICC* [6]. This prompted Fleiss and colleagues to correctly describe these three statistics as belonging to a family of mathematically inter-related coefficients. An analogy in the broader bio-statistical world is the often cited mathematical equivalence between the standard correlation coefficient (r) for interval variables and the Phi coefficient for Nominal-dichotomous variables [17].

The three aforementioned sets of clinical/oenological criteria are given in Table 1.

Table 1. (**A**) The Landis and Koch (1977) Criteria for Assessing Enological Significance; (**B**) The Fleiss (1981) Criteria for Assessing Enological Significance; (**C**) The Cicchetti and Sparrow (1981) Criteria for Assessing Enological Significance.

(**A**)

k, k_w **or** ***ICC***	**Strength of Agreement**
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
>0.80	Almost Perfect

(**B**)

k, k_w **or** ***ICC***	**Clinical Significance**
<0.40	Poor
0.40–0.74	Fair to Good
$\geq$0.75	Excellent

(**C**)

k, k_w **or** ***ICC***	**Clinical Significance**
<0.40	Poor
0.40–0.59	Fair
0.60–0.74	Good
$\geq$0.75	Excellent

In the next section of this report, there will be a discussion of the relevance of a very early, seminal, albeit seldom cited, publication, that nonetheless appears to have made a substantial contribution to our knowledge of how best to understand levels of inter-taster agreement, or agreement more broadly. It recalls in me the musically derived phrase referring to a familiar classic as "an oldie but goodie." This contribution was made by a research Sociologist William Robinson, more than 60 years ago and was published in a prominent research Journal in his field, namely, the *Sociological Review* [18]. Pertinent to this report, Robinson discovered a simple mathematical relationship between what he referred to as the coefficient of agreement (*A*) and the intra-class correlation coefficient (*ICC*). What makes this additional Agreement statistic most desirable, as we will see, is that it is very easy to compute and because of its mathematical relationship to the *ICC*, which is a chance-corrected coefficient, (*A*) itself becomes a chance-corrected coefficient.

5. The Agreement or (*A*) Index and Its Mathematical Relationship to the *ICC*

Suppose two Wine Tasters are asked to rate the quality of each of 200 wines, over a period of one year; and their chance-corrected level of agreement produced an *ICC* value of 0.62 (Good)—[12] or (Substantial)—[8]. If one desires to interpret the *ICC* as another agreement coefficient (*A*), how should one proceed?

The mathematical relationship is given by the very simple formula introduced by Robinson [18] as:

$$\text{Agreement } (A) = (ICC + 1)/2 \tag{1}$$

Given our hypothetical *ICC* value of 0.62, Agreement (*A*) becomes 1.62/2 = 0.81 or 81%.

In Table 2, the author shows the conversion of a given k, k_w or *ICC* value into its Agreement (*A*) equivalent. The relevance this type of thinking has for oenological research is explained in the next section of this report.

There is an additional bio-statistical fact that derives from Robinson's scientific contribution: first, when any of the kappa coefficients is at its highest possible level (Case 1 in each of Tables 4–6), then the level of specific agreement on both Positive and Negative judgments will both be exactly equal to the overall Percentage of Observed agreement (PO).

In the context of clinical research, one earlier investigation had as its focus the accuracy of a number of multiple regression techniques and neural networks (NN) for the binary diagnosis of Autism. Each multiple regression technique (Logistic, Linear and Quadratic) produced more accurate diagnostic results than did Neural Networks. Accuracy was assessed using the standard Sensitivity-Specificity model1 whereby <70% = Poor; 70–79% = Fair; 80–89% = Good; and 90–100% = Excellent [14]. The reader will note that the same set of criteria are used by Robert Parker and other putative experts to evaluate the quality of wine.

Two pertinent questions arise at this point in the narrative: First, what is the correspondence between *ICC* values and Agreement across a broad and comprehensive spectrum of values? Second, how does this information relate to the aforementioned sets of criteria defining levels of enological significance?

The answer to the first query appears in Table 2.

Table 2. The Correspondence between ICC and Percent Agreement [18] [1].

ICC Value	Percent Agreement (*A*)
0.00 (P)	50 (P)
0.05 (P)	52.5 (P)
0.10 (P)	55(P)
0.15 (P)	57.5 (P)
0.20 (P)	60 (P)
0.25 (P)	62.5 (P)
0.30 (P)	65 (P)
0.35 (P)	67.5 (P)
0.40 (F)	70 (F)
0.45 (F)	72.5 (F)
0.50 (F)	75 (F)
0.55 (F)	77.5 (F)
0.60 (G)	80 (G)
0.65 (G)	82.5 (G)
0.70 (G)	85 (G)
0.75 (G)	87.5 (G)
0.80 (E)	90 (E)
0.85 (E)	92.5 (E)
0.90 (E)	95 (E)
0.95 (E)	97.5 (E)
1.00 (E)	100 (E)

[1] Because of the mathematical equivalencies between *ICC*, Kappa and Weighted Kappa, this relationship holds for each of these three statistics for assessing levels of wine tasters' binary judgments, as well as inter-rater agreement levels more generally. See text for more details. The letters P, F, G and E refer, in this context, to Poor, Fair, Good and Excellent wine quality, respectively, as defined by the Robert Parker and similar wine rating scales.

The answer to the second question appears next.

6. Revising the Criteria for the Enological Significance of Research Findings

In order to produce a correspondence between the aforementioned trifecta of clinical significance criteria with the rating of the quality of wine by the Robert Parker or similar scales, a few minor but enologically significant changes need to be made in each of the three sets of guidelines. It should be recalled that the Parker scale for rating the quality of wine is already equivalent to the clinical criteria given by the aforementioned investigation by Cicchetti, et al. [14].

This minor revision process will be illustrated first with the Landis and Koch guidelines [8]. Because of the conceptual similarity between this triad of recommended guidelines, the same logic will apply to the Fleiss guidelines [10] and also those published by Cicchetti [12]. If we now present again the original Landis and Koch guidelines, we have the following illustration:

k, k_w, *ICC*	Agreement (A)	Strength of Agreement
<0.00	<50%	Poor
0.00–0.20	50–60%	Slight
0.21–0.40	60.5–70%	Fair
0.41–0.60	70.5–80%	Moderate
0.61–0.80	80.5–90%	Substantial
0.81–1.00	90.5–100%	Almost Perfect

Note first that the Parker wine quality rating scale, as Percentages, defines below 70 as Poor; 70–79 as Fair; 80–89 as Good and 90 and above as Excellent. In contrast, each of the acceptable-wine-quality scores in the Landis and Koch guidelines appears at the end of each category rather than at its entry level [8]. By simply subtracting the number one from the Slight, Fair, Moderate and Substantial guidelines; and then combining the first two categories Poor and Slight, the revised Agreement categories become 50–69 = Poor; 70–79 = Fair; 80–89 = Good; and 90–100 = Excellent, which, in this revised format, coincides exactly with the clinical criteria for bio-behavioral diagnoses [14], as well as with the Parker quality of wine criteria.

Applying the same logic to the Fleiss, et al. guidelines, the Fair to Good category of k, k_w or *ICC* as 0.40 to 0.74 was changed to 0.40 to 0.79; and the last category was revised to define Excellent at ≥0.80 instead of at ≥0.75.

Finally, the Cicchetti criteria [12] required that the original category of 60 to 74, representing Good Agreement, be revised to 60–79; and that the final category defining Excellent as ≥75 be replaced by ≥80.

These revised criteria, with very minor changes, are now in line with both the aforementioned clinical guidelines [14] and the identical set of Parker criteria for judging the quality of wine. These revised criteria appear in Table 3A–C.

Table 3. (**A**) Revised Landis and Koch Criteria [8] for Assessing Enological Significance; (**B**) Revised Fleiss, Levin and Cho Paik (2003) Criteria for Assessing Enological Significance; (**C**) Revised Cicchetti (1994) Criteria for Assessing Enological Significance.

	(**A**)	
k, k_w or *ICC* **Value**	**Percent Agreement (*A*)**	**Strength of Agreement**
<0.00	<50	Poor
0.00–0.19	50–59.5	Slight
0.20–0.39	60–69.5	Fair
0.40–0.59	70–79.5	Moderate
0.60–0.79	80–89.5	Substantial
>0.80	>90	Almost Perfect
	(**B**)	
k, k_w or *ICC* **Value**	**Percent Agreement (*A*)**	**Clinical Significance**
<0.40	<70	Poor
0.40–0.79	70–89.5	Fair to Good
≥0.80	≥90	Excellent
	(**C**)	
k, k_w or *ICC* **Value**	**Percent Agreement (*A*)**	**Clinical Significance**
<0.40	<70	Poor
0.40–0.59	70–79.5	Fair
0.60–0.79	80–89.5	Good
≥0.80	≥90	Excellent

Thus far, the focus has been on clinical, practical, or, in this context enological significance. This is critical because a research result, enological or otherwise, must have value beyond its level of statistical significance. It must also have clinical, practical or enological significance to be worth pursuing further. Thus, the desideratum must be that a given scientific finding should not only occur beyond chance expectation, it must also not be a trivial finding. For a comprehensive discussion of this fundamental issue, the interested reader is referred to the scholarly work of Borenstein [19].

Thus far, the focus has been on the overall levels of inter-taster agreement or the overall level of chance-corrected agreement, again on an overall level. In the next part of this report, the issue of specific category agreement will be pursued.

7. Specific Category Agreement Levels

In the binary taster agreement context, one is referring to the agreement on positive and negative taster judgments. For example, let us suppose that the enological researcher is investigating the reliability level of inter-taster agreement as to whether wines are oaked (+) or unoaked (−) and the overall agreement, based upon 100 wines, is 80%; she wishes to proceed further and asks the question "What is the agreement on the oaked wines and the unoaked wines, treated separately?" Conceptually, overall agreement, as one might expect, is a weighted average of the agreement on positive and negative cases. In order to explain the phenomenon in greater detail, consider the hypothetical results of an enological wine investigation in which, say, two experienced Wine Tasters are asked to decide whether 100 wines, evaluated over a period of six months, are oaked or not. Suppose the results, in binary contingency table format, are as in this illustration:

	Taster B		
Taster A	Oaked (+)	Unoaked (−)	Totals
Oaked (+)	60	20	80
Unoaked (−)	0	20	20
Total	60	40	100

Summing along the main diagonal, the overall level of Taster agreement is 80%. The agreement on Positive cases is 60/(80 + 60)/2 or 60/70 = 85.7%; this is based on an average of (80 + 60)/2 = 70 cases; the agreement on Negative cases, correspondingly, is 20/(20 + 40)/2 or 20/30 = 66.7%; this derives from an average of (20 + 40)/2 = the remaining 30 cases. Finally: [(85.7 × 70) + (66.7 × 30)] = (60 + 20) = 80%.

8. The Sensitivity-Specificity Model in an Enological Context

The relevance of the Sensitivity-Specificity model for studying the accuracy of Tasters' binary judgments about wine was recently investigated [2]. Given its relevance for this report, it seems pertinent to briefly allude to it once again. The five components of the model have their roots in bio-behavioral diagnostic issues.

The five components of the Sensitivity-Specificity model are Overall Accuracy (OA). This refers to the percentage of correct binary judgments summed over both positive and negative cases. Thus, if there were Taster agreement on 42 of the Positive cases (the wines are oaked) and a corresponding level of agreement on 38 of the Negative cases (the wine is unoaked), the overall agreement level would be 80%. Sensitivity (Se) measures the percentage of oaked wines that are correctly judged as such. If, of 48 wines known to be oaked, 42 were judged correctly by the Tasters, Se would be calculated as 42/48 = 87.5%, Specificity (Sp) would indicate the percentage of unoaked wines that are correctly judged as such. Therefore, if 38 out of 52 wines were judged accurately to be unoaked, Sp would become 38/52 = 73%. Predicted Positive Accuracy (PPA) refers to the percentage of wines that the Tasters judge to be oaked that are actually oaked. Thus, if 42 of 56 wines that are judged to be oaked turn out to indeed be oaked, then PPA would become 42/56 = 75%.

Predicted Negative Accuracy indicates the percentage of wines that the Tasters judge to be unoaked that are actually unoaked. If this were true of 38 of 44 wines, then PPN would become 38/44 = 86%. We now turn to the issue of statistical significance.

9. Criteria for Assessing Levels of Statistical Significance

There are many statistical tests for establishing the level of statistical significance of a given research finding; common among them are the *t* test, the *F* test and the *Z* test, which are all mathematically related to each other.

As pertains to the current investigation, the statistical significance of a given kappa value is found by dividing kappa by its standard error (SE)—[20], which produces a *Z* score, the size of which is directly translated into a probability (*p*) value which is interpreted in the usual way, as: <±1.96 = Not Statistically Significant (NS); ±1.96 = 0.05; ±2.58 = 0.01; ±3 = 0.003; ±4 = <0.0005; and ±5 = <0.0001 [20,21].

Irrespective of which statistic is most appropriate to utilize, the objective is always to determine whether a given research finding (enological or otherwise) has occurred beyond chance expectation. The standard definition of a chance finding is that it must have occurred at or less than five times in 100. Although criticized by some, this "Holy Grail" criterion for statistical significance has withstood the test of time as it continues to be defined at the level of 0.05 probability (*p*).

Given the topic investigated here, the focus will be on binary wine tasting judgments, but for reasons already given, the findings will also apply, conceptually, to other types of variables, ordinal, interval or ratio. In two recent investigations, one clinical, the other oenological, exceedingly high to perfect correlations were found between the reliability and accuracy of binary judgments [1,7].

These results are recast in an enological format at the following levels of overall Wine Taster agreement on a hypothetical binary variable, such as, whether a wine was oaked or not, with overall hypothetical Taster agreement levels set at 70%, (Table 4); at 80% (Table 5) or at 90% (Table 6). In each of these three hypothetical enological data sets, it was possible to control for a number of variables that would be difficult if not impossible to control in the typical enological investigation. These variables were controlled at each hypothetical level of overall Taster agreement, whether 70% (Fair), 80% (Good); or 90% (Excellent), as the following:

For OA = 70%:

The patterns of agreement on Positive and Negative cases were set at 35–35; 40–30; 45–25; 50–20; 55–15; 60–10; 65–5; and 70–0.

The numbers of disagreement cases (+ −) and (− +) were each set at 15. This strategy served two important research purposes: first, to control or eliminate hypothetical Wine Taster bias; more specifically, whenever there was a taster disagreement the first taster was just as likely as the second taster to judge a disagreed upon wine as oaked or unoaked. This same design strategy was utilized for the 80% and 90% condition. It should be noted here that very high levels of inter-taster bias have been demonstrated in the judgments of wine experts such as Jancis Robinson and Robert Parker [22,23].

The outcome variables for each of the 70%, 80% and 90% conditions were the following: The Percentage of agreement expected on the basis of Chance alone (PC); the levels of kappa (*k*) or chance-corrected agreement; the levels of agreement on both Positive and Negative cases, for example, the wine is oaked (+) or the wine is unoaked (−).

The absolute difference between agreement on Positive and Negative wine Tasting judgments, whereby 0 difference = 100% agreement, and maximum possible disagreement would then be 0% agreement; and the final column in each of the three Tables contains the *p* values for each kappa value.

For OA = 80%:

The patterns of agreement on Positive and Negative cases were set at 40–40; 45–35; 50–30; 55–25; 60–20; 65–15; 70–10; 75–5; and 80–0. The numbers of disagreement cases (+ −) and (− +) were each set at 10.

For OA = 90%:

The patterns of agreement on Positive and Negative cases were set at 45–45; 50–40; 55–35; 60–30; 65–25; 70–20; 75–15; 80–10; 85–5; and 90–0. The numbers of disagreement cases (+ −) and (− +) were each set at 5.

The advantage of the hypothetical information revealed in these three tables is that they allow for a degree of experimental control that is seldom or almost never possible in the typical enological study or in clinical research more generally. The method of Hypothetics, or Enothetics in this context, allows the research scientist to produce the results that would have occurred if the actual experiments they represent were feasible. The general findings will precede those occurring on a case by case basis, separately for the 70%, 80% and 90% condition.

Table 4. Relationship between the Reliability and Accuracy of Pairs of Hypothetical Tasters Judging Whether a Wine is Oaked (+) or Unoaked (−) When the Tasters Are in 70% Agreement.

Case:	(++)	(− −)	(+ −)	(− +)	PC	Kappa [1]	PO^+	PO^-	PO^+/PO^- Agreement	*p* Value [3]
1	35	35	15	15	50	0.40 (F)	70 (F)	70 (F)	100	<0.0005
2	40	30	15	15	50.5	0.39 (P)	67 (P)	73 (F)	94	0.002
3	45	25	15	15	52	0.375 (P)	62.5 (P)	75 (F)	87.5	0.004
4	50	20	15	15	54.5	0.34 (P)	57 (P)	77 (F)	80	0.01
5	55	15	15	15	58	0.29 (P)	50 (P)	79 (F)	71	NS [2]
6	60	10	15	15	62.5	0.20 (P)	40 (P)	80 (G)	60	NS
7	65	5	15	15	68	0.06 (P)	25 (P)	81 (G)	44	NS
8	70	0	15	15	74.5	−0.18 (P)	0 (P)	82 (G)	18	NS

The correlation between the size of kappa and the difference in agreement on Positive and Negative cases is +0.98; [1] Kappa values are classified as Poor (P), Fair (F), Good (G) or Excellent (E) by the revised Cicchetti criteria in Table 3C; [2] NS = not statistically significant at $p \leq 0.05$; [3] Statistical significance is found by dividing kappa by its standard error as derived by Fleiss, Cohen and Everitt, [20]. Values of Z are interpreted in the standard manner whereby <±1.96 = *p* at the 0.05 level; ±2.58 is at t 0.01; ±3 at 0.003; ±4 at 0.0005; and ±5 at 0.0001 [20,21].

Table 5. Relationship between the Reliability and Accuracy of Pairs of Hypothetical Tasters Judging Whether a Wine is Oaked (+) or Unoaked (−) When the Tasters Are in 80% Agreement.

Case:	(++)	(− −)	(+ −)	(− +)	PC	Kappa [1]	PO^+	PO^-	Agreement	PO^+/PO^- *p* Value [3]
1	40	40	10	10	50	0.60 (G)	80 (G)	80 (G)	100	<0.0005
2	45	35	10	10	50.5	0.60 (G)	82 (G)	78 (F)	96	<0.0005
3	50	30	10	10	52	0.58 (F)	83 (G)	75 (F)	92	0.001
4	55	25	10	10	55	0.56 (F)	85 (G)	71 (F)	86	<0.005
5	60	20	10	10	58	0.52 (F)	86 (G)	67 (P)	81	<0.005
6	65	15	10	10	63	0.47 (F)	87 (G)	60 (P)	73	<0.005
7	70	10	10	10	68	0.38 (P)	88 (G)	50 (P)	62	0.01
8	75	5	10	10	74.5	0.22 (P)	88 (G)	33 (P)	45	NS [2]
9	80	0	10	10	82	−0.11 (P)	89 (G)	0 (P)	11	NS

The correlation between the size of kappa and the difference in agreement on Positive and Negative cases is +0.99; [1] Kappa values are classified as Poor (P), Fair (F), Good (G) or Excellent (E) by the revised Cicchetti criteria in Table 3C; [2] NS = not statistically significant at $p \leq 0.05$; [3] Statistical significance is found by dividing kappa by its standard error as derived by Fleiss, Cohen and Everitt, (1969). Values of Z are interpreted in the standard manner whereby <±1.96 = *p* at the 0.05 level; ±2.58 is at *t* 0.01; ±3 at 0.003; ±4 at 0.0005; and ±5 at 0.0001 [20,21].

Table 6. Relationship between the Reliability and Accuracy of Pairs of Hypothetical Tasters Judging Whether a Wine is Filtered (+) or Not Filtered (−) When the Tasters Are in 90% Agreement.

Case:	(++)	(− −)	(+ −)	(− +)	PC	Kappa [1]	PO^+	PO^-	(PO^+/PO^-)	*p* Value [3]
1	45	45	5	5	50	0.80 (E)	90	90	100	<0.0005
2	50	40	5	5	51	0.80 (E)	89	91	98	<0.0005
3	55	35	5	5	52	0.79 (G)	88	92	96	<0.0005
4	60	30	5	5	55	0.78 (G)	86	92	94	<0.0005
5	65	25	5	5	58	0.76 (G)	83	93	90	<0.0005
6	70	20	5	5	63	0.73 (G)	80	93	87	<0.0005
7	75	15	5	5	68	0.69 (G)	75	94	81	<0.0005
8	80	10	5	5	75	0.61 (G)	67	94	73	<0.0005
9	85	5	5	5	82	0.44 (F)	50	94	56	0.001
10	90	0	5	5	90.5	−0.05 (P)	0	95	5	NS [2]

The correlation between the size of kappa and the difference in agreement on Positive and Negative cases is +1.00; [1] Kappa values are classified as Poor (P), Fair (F), Good (G) or Excellent (E) by the revised Cicchetti criteria in Table 3C; [2] NS = not statistically significant at $p \leq 0.05$; [3] Statistical significance is found by dividing kappa by its standard error as derived by Fleiss, Cohen and Everitt [20]. Values of Z are interpreted in the standard manner whereby <±1.96 = *p* at the 0.05 level; ±2.58 is at *t* 0.01; ±3 at 0.003; ±4 at 0.0005; and ±5 at 0.0001 [20,21].

10. Overall Results: Correlations between the Reliability and Accuracy of Wine Tasters' Hypothetical Binary Judgments

As we examine the results deriving from Tables 4–6, it should be noted that the correlations between reliability and overall accuracy or validity of Tasters' hypothetical binary wine judgments is exceptionally high, that is, almost perfect to completely perfect. This holds true whether the overall Taster agreement levels were expressed at 70% (Fair); 80% (Good) or 90% (Excellent). The three correlations are, respectively, +0.98, +0.99 and +1.00.

An advantage of using the standard correlation coefficient to measure the relationship between the reliability and accuracy/validity of hypothetical Tasters' judgments is that it provides a familiar and easy-to-interpret result. A major disadvantage is that the correlation coefficient is an omnibus statistic that provides no information about reliability and accuracy of judgment on a case by case basis, as would be true of individual *kappa* coefficients or the components of the Sensitivity-Specificity model in whatever clinical or other research context.

11. Hypothetical Results on a Case by Case Basis

With respect to the hypothetical data in Table 4 (patterns of 70% agreement), the Case 1 result indicates both enological and statistical significance; and Cases 5 through 8 indicate results that are neither enologically nor statistically significant; however, Cases 2–4 produce findings that are statistically significant but not enologically significant.

The results for the 80% condition, as spread in Table 5, show that the first six Cases yield results that are both enologically and statistically significant, while the seventh Case indicates a result that is statistically significant but not enologically significant; while Cases 8 and 9 are neither enologically nor statistically significant.

The data for the 90% condition indicates that the first nine Cases produced results that are both enologically and statistically significant while the result for the tenth Case was neither enologically nor statistically significant.

Taken as a whole, these results are consonant with two research results that occur in both the enological and clinical world of science: first, given an appropriate sample size, even the most trivial of results will be statistically significant; and secondly, the greater the level of agreement, the more likely the result is apt to be both statistically significant and of material importance, whether enological, clinical, or otherwise.

One way to provide more specific information on a case by case basis is to summarize the data from Tables 4–6 in a single table as follows: The hypothetical information for the 70% condition was based upon eight cases; the 80% condition on nine cases; and the 90% condition was based upon an additional 10 cases. These sum to 27 cases in all.

If one now recasts the data into a 2 × 2 or binary Table, it will then be possible to perform the *kappa* statistic, to measure the level of hypothetical Taster reliability as well as to obtain the five accuracy components of the Sensitivity-Specificity model. The recast data appear in Table 7.

Table 7. Illustrating the Relationship between the Reliability and Accuracy of Wine Tasters' Hypothetical Binary Judgments of Whether a Wine is Oaked (+) or Not Oaked (−), Expressed in Percentages.

	Taster 2		
Taster 1	**(+)**	**(−)**	**Totals**
(1)	12	4	16
(−)	0	11	11
Totals	12	15	27

The summary data in Table 7 indicate the following:

Overall Agreement = 23/27 = 85.2%. Chance Agreement = 13.2/27 = 48.9%.
Kappa = (85.2 − 48.9)/51.1 = 0.71; p = 0.001.
These results, from an oenological viewpoint, are: Substantial by the Landis & Koch criteria; acceptable by the Fleiss, et al. criteria; and Good by the Cicchetti criteria.
Se = 12/12 = 100% (Perfect)
Sp = 11/13 = 85% (Good)
PPA = 12/16 = 80% (Good) and
PNA = 11/11 = 100% (Perfect)

12. Summary and Conclusions

Utilizing a new methodology called Hypothetics, or Enothetics in a wine tasting investigation, a model was introduced that makes it possible to control for a large number of variables that are often most difficult to control in the typical enological study, beverage study or more generally. Because of this level of control, the method allows for findings and insights that are often not possible using available standard methodologies and standard data analytic strategies. In this fundamental sense, the hypothetical results that were obtained appear to have heuristic value for the design of future enological studies and investigations focusing on beverages more generally.

In this application, which focused upon the enological and statistical significance of Wine Tasters' binary judgments, the following occurred: Results that were enologically significant (had practical meaning) were uniformly statistically significant, although the reverse was not always true; that is to say, a number of results were statistically significant, but not enologically important. A method developed more than six decades ago [19] was shown to simplify the understanding of chance corrected agreement coefficients in any given enological or other type of scientific investigation. Finally, the correlation between the reliability and validity or accuracy of binary judgments was shown to be exceedingly high whether on an overall omnibus level; or on a case by case basis. With appropriate adjustments, this methodology would apply to ordinal and interval variables, as well.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cicchetti, D.V. Opinions versus facts: A bio-statistical paradigm shift in oenological research. *Proc. J. Wine Res.* **2017**, *1*, 1–8.
2. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *23*, 37–40. [CrossRef]
3. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 195–201. [CrossRef]
4. Bartko, J.J. The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* **1966**, *19*, 3–11. [CrossRef] [PubMed]
5. Bartko, J.J. Corrective note to "The intraclass correlation coefficient as a measure of reliability". *Psychol. Rep.* **1974**, *34*, 1–11. [CrossRef]
6. Shrout, P.E.; Fleiss, J. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [CrossRef] [PubMed]
7. Cicchetti, D.V.; Klin, A.; Volkmar, F.R. Assessing binary diagnoses of bio-behavioral disorders: The clinical relevance of Cohen's Kappa. *J. Nerv. Ment. Dis.* **2017**, *205*, 58–65. [CrossRef] [PubMed]
8. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *3*, 159–174. [CrossRef]
9. Fleiss, J. *Statistical Methods for Rates and Proportions*, 2nd ed.; Wiley: New York, NY, USA, 1981.
10. Fleiss, J.; Levin, B.; Paik, M.C. *Statistical Methods for Rates and Proportions*, 3rd ed.; Wiley: New York, NY, USA, 2003.
11. Cicchetti, D.V.; Sparrow, S.S. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *Am. J. Ment. Defic.* **1981**, *86*, 127–137. [PubMed]
12. Cicchetti, D.V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* **1994**, *6*, 284–290. [CrossRef]

13. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Erlbaum: Hillsdale, NJ, USA, 1988.
14. Cicchetti, D.V.; Volkmar, F.R.; Klin, A.; Showalter, D. Diagnosing autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. *Child Neuropsychol.* **1995**, *1*, 26–37. [CrossRef]
15. Fleiss, J. Measuring agreement between two judges on the resence or absence of a trait. *Biometrics* **1975**, *31*, 651–659. [CrossRef] [PubMed]
16. Fleiss, J.L.; Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **1973**, *33*, 613–619. [CrossRef]
17. Kaltenhauser, J.; Lee, Y. Correlation coefficients for binary data. *Geogr. Anal.* **2010**, *8*, 305–313. [CrossRef]
18. Robinson, W. The statistical measurement of agreement. *Am. Sociol. Rev.* **1957**, *22*, 17–25. [CrossRef]
19. Borenstein, M. The shift from significance testing to effect size estimation. *Res. Methods Compr. Clin. Psychol.* **1998**, *3*, 319–349.
20. Fleiss, J.L.; Cohen, J.; Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **1969**, *72*, 323–327. [CrossRef]
21. Cohen, J.; Cohen, P.; West, S.G.; Aiken, I.S. *Appliede Multiple Regression/Correlation for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum: Mahwah, NJ, USA, 2003.
22. Cicchetti, D.V.; Cicchetti, A.F. As wine experts disagree, consumers' taste buds flourish: How two experts rate the 2004 Bordeaux vintage. *J. Wine Res.* **2013**, *24*, 311–317. [CrossRef]
23. Cicchetti, D.V.; Cicchetti, A.F. Two enological titans rate the 2009 Bordeaux wines. *Wine Econ. Policy* **2014**, *3*, 28–36. [CrossRef]