Wine rating scales: Assessing their utility for producers, consumers, and oenologic researchers

Domenic V Cicchetti[1]
Arnold F Cicchetti[2]

[1]Yale Home Office, North Branford, CT, USA; [2]Consulting Specialist of Wine, Director of National Accounts, Wine Warehouse, San Anselmo, CA, USA

Abstract: The authors studied seven wine rating scales judged to be useful for the wine producer, consumer, or oenologic researcher: (1) My Wine Rating scale; (2) the Amerine and Roessler (1983) wine rating system; (3) the redwinebuzz.com rating system; (4) Robert Parker's wine rating scale; (5) the *Wine Spectator* scale; (6) the Stephen Tanzer scale; and (7) the Chebnikowski Winespider evaluation system. Statistics were applied to answer three hypothetical oenologic questions: (1) Does a particular wine meet criterion for everyday consumption? (2) How well do tasters agree on both the level of wine aroma and bouquet? And (3) How well do tasters agree on the overall quality of a wine, both currently, and after it has been appropriately cellared for 10 years? The implications of this study were discussed in terms of their heuristic value for further oenologic research. One fundamental issue that has received little or no attention in oenologic research is a determination of measurable taster variability that can be expected to occur when the same wine is blind tasted again, either by the same taster or by other tasters blindly evaluating the same wine.
Keywords: wine, rating scales, taster consistency

Introduction

The objectives of this article are to: (1) provide an overview of the major extant wine rating scales; (2) describe their structure or anatomy; and (3) Give hypothetical examples of how their reliability would be assessed, and the information that would thereby be obtained.

An overview and description of major wine rating scales

A search of the oenologic literature reveals a number of wine rating scales that differ along a range of interesting characteristics, namely, the number of rating categories; the number and type of wine descriptors, and the methods by which they are scored and interpreted. In order to be selected, a given wine scale had to meet the following objective requirements: (1) Each category of wine evaluation needed to provide distinctive and distinguishing information pertaining to level of wine quality; and (2) The wine scale had to be designated as one that could be used meaningfully by one or more of the following: the wine producer, the wine consumer, or the oenologic researcher.

Application of these two criteria resulted in selecting the following wine rating scales: (1) the 10-point My Wine Rating Scale; (2) the Amerine and Roessler scale,[1] developed at UC Davis, in 1983; (3) the redbuzz.com wine rating scales and the 100-point wine rating scales of (4) Robert Parker (the Wine Advocate); (5) Steven Tanszer; (6) the *Wine Spectator* staff, and (7) Nick Chebnikowski's Winespider evaluation system, the latter a 160-point scale, whose total scores are easily converted to a 100-point scale.

The structure or anatomy of each of these wine rating scales, as well as their utility for the producer, consumer, and oenologic researcher, will be elucidated in the next section.

Correspondence: Domenic V Cicchetti
Yale Home Office, 94 Linsley Lake Road, North Branford, CT, USA
Tel +1 203 488 6563
Fax +1 203 483 1123
Email dom.cicchetti@yale.edu

© 2009 Cicchetti and Cicchetti, publisher and licensee Dove Medical Press Ltd. This is an Open Access article which permits unrestricted noncommercial use, provided the original work is properly cited.

Major wine rating scales: Structure/anatomy

My wine rating scale

Developed by "Tim" (March 9th, 2005), this 7-point scale uses a "Numerical rank" defined by descriptors that the author refers to as "Word ranks." Each Word rank descriptor is defined somewhat obliquely by their respective author "Comments". Table 1 shows the My Wine Rating scale.

This scale can be utilized as either a full 10-point or 7-point wine rating scale (if one chooses to collapse ratings of <4 into a single oenologic category, "Undrinkable"). Because of the colloquial manner in which the comments are addressed, the scale would be of lesser use to either the wine producer or the oenologic researcher. However, the scale does have the advantage of providing clear and nonoverlapping descriptors to distinguish one "Word rank" from another. As such, it could be used by experienced or highly sophisticated wine tasters. One would probably rule out its use by the average or neophyte wine imbiber. For more details, the interested reader is referred to Winecast (see http://winecast.net/2005/03/09/my-wine-ratingscale/).

The Amerine and Roessler (1983) wine rating system[1]

This wine rating system allows for: (1) an overall, or global assessment of a wine's quality, on a 21-point scale, ranging between 0 and 20 (with higher scores denoting better wine quality); and (2) a rating or evaluation of specific characteristics of the wine, as measured on separate rank-ordered, or ordinal subscales, and also constructed so that the higher the score, the better the quality. These include ratings of: appearance and color (0–2), aroma and bouquet (0–6), total acidity (0–1), balance (0–2), body (0–1), flavor (0–3), and finish (0–2).

As an example, the wine characteristic, "aroma and bouquet" is scored according to one of the following six characteristics, with its separate, nonoverlapping criteria:

6. Extraordinary. Unmistakable characteristic aroma of grape variety or wine type. Outstanding and complex bouquet. Exceptional balance or aroma bouquet
5. Very good. Characteristic aroma. Complex bouquet. Well balanced.
4. Good. Characteristic aroma. Distinguishable bouquet.
3. Pleasant. Slight aroma and bouquet, but pleasant.
2. Acceptable. No perceptible aroma or bouquet.
1. Objectionable. Objectionable with off odors.

The redwinebuzz.com rating system

The current scale consists of 5-point ordinal or rank ordered subscales for rating the following wine characteristics: color, nose, palate, finish, tannins, acidity, alcohol, aging potential, and food friendliness. In addition, the rating system provides for a 10-point evaluation of the overall quality of a given wine. According to the authors of the scale, this latter category will be rated on a 6-point ordinal scale in the revised version of this rating scale. Each ordinal category contains very carefully thought–through criteria for defining the various levels of wine characteristics. As an example, the five category ordinal scale that reflects the "complexity, duration, and harmony" of a given red wine varietal has the following criteria for evaluating where the wine fits on this scale:

5. "Very complex and persistent flavors"
4. "Complex flavors"
3. "Medium complexity of flavors"
2. "Straightforward flavors"
1. "Very simple, vague flavors"

This scale is quite sophisticated in its psychometric structure, with carefully considered, quite comprehensive, nonoverlapping descriptors for each of the aforementioned wine characteristics. The scale also uses descriptors that are specific to red and white varietals. As a prototypic example, Table 2 shows the scoring for "Palate."

Similarly, the perceived "Overall quality" of a given wine in the new version of this wine rating scale, will be scored as seen in Table 3.

This carefully crafted wine scale would have appeal and be quite useful for the wine producer, the consumer, and the oenologic research scientist. More detailed information about this wine rating scale is available at redwinebuzz.com (see http://www.redwinebuzz.com/).

The next three wine scale rating scales share a number of characteristics: (1) They were each constructed by professional wine evaluators (Robert Parker, of the Wine Advocate; Stephen Tanzer; and professional wine tasters on the *Wine Spectator* staff); (2) They each consist of a percentage score that categorizes wine ratings into excellent or outstanding, very good, good, fair, or poor (not recommended); and (3) They provide an overall wine rating, with no attention paid (except in the summary tasting notes) of specific wine characteristics, such as palate, acidity, balance, finish, etc.

The Parker wine rating scale

The Parker wine rating scale (Table 4) would be of interest to wine producers, wine consumers, and oenologic researchers. More detailed information about this scale is available

Table 1 My wine rating scale

Numerical rank	Word rank	Comments
10	Excellent	Heitz "Martha's Vineyard" Cabernet comes to mind as the best wine I have ever had the good fortune to taste and would earn a "10". This rating is reserved for classic wines.
9	Delicious	A wine of complexity and distinction; the top end of wines tasted on the show to date.
8	Very good	Highly recommended wine; one that I probably have in my cellar right now.
7	"Quaffable"	Like Miles says in *Sideways*, a well made, but ultimately nondistinctive wine. Nothing wrong here, just not a transcendent wine experience.
6	Fair, but has noticeable flaws	A drinkable wine, but one that is not recommended due to wine-making problems or thin fruit flavors.
5	Pretty bad	A wine that is on the verge of being undrinkable; avoid!
1–4	Undrinkable	I would demand a refund should I have the bad fortune to taste a wine that rates a 4 or below.

Table 2 Redwinebuzz.com: Palate score

Score	Wine characteristics
5	Very complex and balanced. This score represents most of the full spectrum of possible flavors expected of this type of wine (at least 4). These flavors show diversity and depth. They are complex and multilayered, showcasing the wine's full potential. Unexpected but pleasant and complementary flavors are possibly present. Great harmony and balance are required for this score.
4	Wines with this score display pleasant flavors most typical of this style but not the rarer components of the flavor of an exceptional wine (typically 3). There is good complexity and depth but possibly to a lesser degree than the higher score category. Balance and harmony are not ideal and alcohol may be excessive.
3	Wines with this score display pleasant flavors most typical of this style that are average to just above average in complexity. These wines display an even narrower spectrum and lesser complexity of expected flavors than the next higher score category (usually 2–3). The flavors may be straightforward. Balance may be problematic and alcohol may be excessive.
2	Complexity of flavors in wines with this score is in the average range (usually no more than 2). Flavors are those typically expected of the varietal but are not evocative or captivating. Wine may seem inordinately thin or lean and balance may be problematic giving the wine a crudely constructed character.
1	A sub-par wine lacking pleasant fruit or wine style-related flavors. Only one identifiable flavor is present. May also have off-putting or unpleasant flavors. Flawed. TCA taint is not included.

Table 3 Overall quality score

Score	Characteristics
5	Outstanding, exceptional, rare and classic. Must-have. This is a superb wine exemplifying the best this style and region has to offer. This offering exceeds the wine maker's stylistic intentions.
4	Very good. Highly recommended. Very good wine reflective of high standards of this style and region. Meets or exceeds wine maker's vision.
3	Good, with appealing characteristics. Recommended. A good wine. Characteristics are in the upper ranks of wines in this style and from this region. Meets or falls short of the wine maker's stylistic intentions for a good, approachable wine.
2	Fairly good. No serious flaws. Worth trying. Demonstrates the general traits of a wine of this style and region. Meets or falls short of the wine maker's stylistic intentions for a good, approachable wine.
1	Average quality. May have prominent flaws. May appeal to some. Not very complex. Falls short of aspirations to a higher standard.
0	We are unable to recommend the wine in question.

Table 4 The Parker wine rating scale

Wine rating	Wine description
96–100	An extraordinary wine of profound and complex character displaying all the attributes expected of a classic wine of its variety.
90–95	An outstanding wine of exceptional complexity and character. In short, these are terrific wines.
80–89*	A barely above average to very good wine displaying varying degrees of finesse and flavor as well as character with no noticeable flaws.
70–79	An average wine with little distinction except that it is soundly made. In essence, a straightforward, innocuous wine.
60–69	A below average wine containing noticeable deficiencies, such as excessive acidity and/or tannin, an absence of flavor, or possibly dirty aromas or flavors.
50–59	A wine deemed to be unacceptable.

Note: Parker does say that wines in the 85–89 range are very very good wines, are often good buys, and that he has many of them in his cellar.

at Robert Parker's website (see http://erobertparker.com/info/legend.asp).

The *Wine Spectator* rating scale

The *Wine Spectator* wine rating scale (Table 5) would also be applicable to consumers, producers and oenologic researchers (see http://www.gotastewine.com/articles/wine-rating-scale.htm).

The Stephen Tanzer wine rating scale

As is true of the Parker and *Wine Spectator* rating scales, the Tanzer scale (Table 6) would be of interest to consumers, producers, and oenologic research scientists. The Tanzer wine rating scale is available from Stephen Tanzer's website (see http://www.wineaccess.com/expert/tanzer/ratingscale.html).

The next and final scale to be discussed is the comprehensive Winespider evaluation system.

The Winespider evaluation system

This wine rating system was created by Nick Chebnikowski, an Australian artist of repute, who also developed a keen interest in matters oenologic (see http://www.nicks.com.au/Index.aspx?link_id=77.459\).

The Winespider evaluation system is more comprehensive than any of the aforementioned scales in that it is comprised of 16 wine attributes, each measured on a 10-point ordinal scale, specifically: color, viscosity, brilliance, depth, aroma, faults, varietal, intensity, complexity, concentration, fruit, length, aftertaste, balance, tannins, and acid. The total maximum or overall wine rating score, then becomes $(10 \times 16) = 160$. As will be illustrated, any given overall wine score is easily converted to a score that can range, potentially, between 16 and 100.

For example, let us assume (see http://www.nicks.com.au/Index.aspx?link_id=77.459) that a particular varietal, such as pinot noir, receives perfect scores of 10 each for eight wine characteristics: balance, tannins, acid, color, viscosity, brilliance, depth, and faults. For the remaining eight wine characteristics, aftertaste, length, fruit, concentration, complexity, intensity, varietal, and aroma, it receives scores of 7, 8, 9, 9, 7, 7, 8, 8, respectively.

The total score would be $[(10 \times 8) + 63]/160$, that converts to 89.4 or 89%. The wine's eponymously implied "spider-web" profile would depict a spider-web visual pattern (Figure 1).

It should be noted that the Winespider evaluation system is unique among wine rating scales in that it can also be used to track whatever changes may occur in any of the 16 wine characteristics, as the wine, a living organism, changes over time. However, as noted correctly on the website, the accuracy of the follow-up wine ratings will depend upon the quality of care the cellared wine receives from the consumer. Also, with a few notable exceptions, it is better to drink wines when they are younger, rather than older, in order to enjoy the optimal levels of their wine characteristics.

Table 5 The wine spectator rating scale

Wine rating	Wine description
95–100	Classic: a great wine.
90–94	Outstanding: wine with superior character and style.
85–89	Very good: wine with special qualities.
80–84	Good: a solid, well-made wine.
70–79	Average: drinkable wines that may have minor flaws.
60–69	Below average: drinkable wine but not recommended.
50–59	Poor: undrinkable wine, not recommended.

Table 6 The Stephen Tanzer wine rating scale

Wine rating	Wine description
95–100	Extraordinary
90–94	Outstanding
85–89	Very good to excellent
80–84	Good
75–79	Average
70–74	Below average
<70	Avoid

This comprehensive wine rating system can be used effectively by wine producers, wine consumers, and oenologic researchers alike. The wine rating system is owned by Nicks Wine Merchants: Vintage Direct (see http://www.nicks.com.au/index.aspx?link_id=77.459; http://www.nicks.com.au/; and the recently updated website, http://www.winespider.com/for more detailed information).

We will now focus upon how wine rating scales can be used to evaluate inter-taster reliability or consistency. In the next section we will describe briefly the statistics of choice. This will be followed by a section showing how the statistics can be applied, with hypothetical oenologic data sets. Information will also be provided concerning the availability of the software required to make the reliability assessments.

However, we need to provide a caveat concerning the application of the 100-point scales just discussed. One major issue that the producers of these scales seem to have ignored is the fact that no clinical instrument is without test–retest variability. This statement of biostatistical fact is not specific to wine scales, but is true of every instrument that is used to make an important clinical judgment. In the world of medicine, this would include basic measures of blood pressure, blood cholesterol levels, and the results of any other laboratory test. This is also true of any psychological measurement, such as the standard IQ test, where it has been demonstrated that one can expect, even in the hands of the very best clinicians, applying the most cutting edge cognitive instruments, a test-retest measurement error of ±5–6 IQ points (eg, Kaufman[2]). Inter-examiner reliability levels can be expected to be even greater, especially if the examiners differ in their level of test expertise.

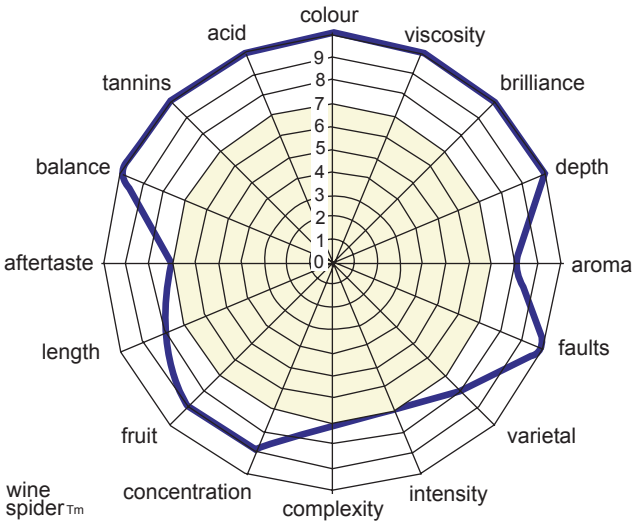

Figure 1 Winespider evaluation system diagram. Copyright © 2009. Vintage Direct. http://www.nicks.com.au/Index.aspx?link_id=77.459.

A mini primer about assessing inter-taster reliability

Defining scales of measurement for oenologic research

Data in general and oenologic data in particular are measured on one of the following types of scales: nominal, ordinal, or continuous. Oenologically speaking, nominal scales consist of two or more categories of wine classification. As the simplest example of a nominal wine rating scale, the consumer decides on the basis of a number of judgments, whether a wine is: (1) worth purchasing; or (b) not worth purchasing.

Oenologic variables can also be measured on an ordinal scale, one that is based upon a rank-ordering of categories. An example would be a rating of the overall quality of a wine as one of the following: poor, fair, good, or excellent. Finally, if Parker's rating scale were employed, the overall quality of a given wine would be a percentage score ranging between 50 and 100. This scoring system would define the scale of measurement as continuous. The work of Cicchetti and colleagues[3] indicated that ordinal scales of measurement containing seven or more categories are as reliable as continuous scales, so that for all intents and purposes ordinal scales of this length can be treated statistically as continuous scales of measurement. This work was confirmed, some years later, by Preston and Colman.[4]

Appropriate statistics for assessing inter-taster agreement or reliability

The type of statistic required to assess the level of agreement or reliability of two or more tasters evaluating a given wine is highly dependent upon the type of scale upon which a given wine rating is made. For simplicity's sake, the statistics of choice, fortunately, are all part of a kappa or kappa-related type reliability statistic.

The statistic, Kappa (κ), was developed by Cohen,[5] and corrected by Fleiss and colleagues[6] and is applicable for data deriving from a nominal category wine rating scale, such as the measurement of whether to purchase a given wine, categorized as yes, no, or undecided. Note that these three categories of classification bear no ordinal or rank–ordered relationship to each other.

The statistic, Weighted Kappa (κ_w), was developed by Cohen[7] and corrected by Fleiss and colleagues.[6] This statistic is applicable to wine scales containing three or more ordinal, or rank–ordered categories, such as the overall rating of a wine as one of the following: outstanding, excellent, good, fair, or unacceptable.

Finally, the statistic called the intraclass correlation coefficient (ICC), was developed by Bartko[8,9] and would be applicable when the wine is measured on a continuous scale. The ICC would be applicable when the overall quality of a given wine was measured using, say, the Parker, *Wine Spectator*, Tanzer, or Winespider wine scoring systems.

κ or κ_w is defined as a ratio in which the difference between observed (PO) and expected (PC) inter-taster agreement (PO–PC) is divided by (1–PC), producing the following formula:

$$\kappa \text{ or } (\kappa_w) = (PO–PC)/(1–PC). \quad (1)$$

Fleiss[10] showed under what conditions κ = ICC; and Fleiss and Cohen[11] showed under what conditions κ_w and the ICC are mathematically identical. These relationships define the three statistics as inter–related or as a family of kappa-type reliability statistics.

κ and κ_w will equal zero when the level of inter-taster agreement (PO) is no better than the agreement expected by chance alone (PC). In such a case, (PO–PC) = 0. When PO > PC (the usual case), then κ or κ_w will assume a positive value; and when there is less inter-taster agreement than one would expect on the basis of chance alone, then κ or κ_w will assume a negative value.

How κ, κ_w, and the ICC are interpreted for levels of statistical significance

In order to test for statistical significance, the value of κ or κ_w is divided by the standard error of κ, forming a Z statistic that is evaluated for level of statistical significance in the usual manner (Table 7).

The level of statistical significance of a given ICC value is based upon a formula deriving from the results of a number of tasters by number of wines rated analysis of variance (ANOVA), and an example of how the statistic may be applied will be illustrated in a later section of this report.

Table 7 Values of Z and level of statistical significance, *P*

Z of κ/κ_w	*p* Value
<1.645	Not significant
±1.645	0.10
±1.96	0.05
±2.57	0.01
±3.00	0.001
+≥4.00	<0.0001

Differentiating statistical significance from clinical significance

As Fleiss[12] correctly noted (and as applied to inter-taster evaluations), in any reliability research design, a certain amount of measurable agreement will occur by chance alone. Accordingly, κ, κ_w, and the ICC are statistics that are defined by a ratio in which the difference between observed and chance agreement form the numerator; and (1-chance) form the denominator. Landis and Koch[13] and Fleiss[12] also went on to say that when the number of raters (read wine tasters) is large, then even a very low κ, κ_w, or ICC value, such as, say, 0.10, will be statistically significant, though clinically meaningless.

As a result, several sets of levels of clinical significance, strength of agreement, or similar terms, such as practical significance have been developed by biostatisticians (Landis and Koch,[13] Fleiss,[12] and Cicchetti and Sparrow).[14] These sets are shown in Tables 8–10.

The reader will note that the three sets of criteria are conceptually quite similar and also that the Landis and Koch[13] criteria are more finely gradated and might therefore be more appropriate in the training of neophyte wine tasters to begin to agree with their more sophisticated imbibers over a period of training sessions.

In the next section we will apply the various κ, κ_w, and ICC approaches to hypothetical tasters who rate independently, a number of wine characteristics: (a) whether a wine is acceptable

Table 8 The criteria of Landis and Koch[13]

Size of reliability coefficient	Strength of agreement
<0.0	Poor
0.0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Table 9 The criteria of Fleiss[12]

Size of reliability coefficient	Clinical significance
<0.40	Poor
0.40–0.74	Fair to good
0.75–1.00	Excellent

for everyday imbibing or not (the nominal case); (b) whether tasters can agree on a wine's level of aroma and bouquet; and whether tasters can agree adequately on the overall quality of a wine, both initially (c); and (d) 10 years later.

Results will be presented in a conceptual, minimally mathematical framework. However, references will be provided that present the formulae, for the more mathematically curious reader.

Which wines are acceptable for everyday imbibing?

Here we will assume that two experienced wine tasters, over a protracted period of time, taste 100 different wines, and score each one as to whether or not it is deemed acceptable for everyday imbibing. Table 11 presents our data.

This fourfold or contingency table format of the data allows the interested oenologic researcher to identify, at a glance, both the 85 wines the two tasters agreed were not acceptable for everyday imbibing as well as the 15 wines upon which the wine tasters disagreed.

Using a computer program in Cicchetti and Heavens[15] and earlier in Heavens and Cicchetti,[16] we obtained the following outputs:

The Proportion of Observed agreement (PO) = 85/100 = 0.85

The Proportion of Chance agreement (PC) = $[(0.60 \times 0.55) + (0.40 \times 0.45)]$
$= (0.33 + 0.18)$
$= 0.51$

$$\kappa = (PO–PC)/(1–PC) \quad (1)$$
$$= (0.85–0.51)/(1–0.51)$$
$$= 0.34/0.49$$
$$= 0.6939$$

Table 10 The criteria of Cicchetti and Sparrow[14]

Size of reliability coefficient	Clinical significance
<0.40	Poor
0.40–0.59	Fair
0.60–0.74	Good
0.75–1.00	Excellent

The level of statistical significance is determined by dividing the size of κ (0.6939), by its standard error (here, a value of 0.0998), to obtain the following values of Z and level of statistical significance, p:

$$Z = K/S.E._k \quad (2)$$
$$= 0.6939/0.0995$$
$$= 6.97, \text{ with } p < 0.0001 \text{ (Table 7).}$$

In terms of level of clinical significance, the κ of 0.69 is considered good, by the aforementioned criteria of Cicchetti and Sparrow[14] and substantial by the criteria of Landis and Koch.[13]

How well do tasters agree on wine aroma and bouquet?

In this hypothetical example, let us assume that two tasters evaluate, independently, and over a suitable period of time, 100 wines, as to level of aroma and bouquet, using the Amerine and Roessler (1983) wine rating system.[1] Recall (see above, Appropriate statistics for assessing inter-taster agreement or reliability) that the system allows the taster to choose one of the 6–1 levels of aroma and bouquet for any given wine.

Assume further that the data display themselves as in Table 12.

Unlike data derived from a nominal scale that can only be scored as "present" or "absent," data deriving from an ordinal or rank–ordered scale when used, say, by two or more wine tasters, can be scored as "present", meriting a perfect score of 1, or 100%; or a score of 0, when the raters are maximally apart (such as 1–6 or 6–1 ratings on a 6-point ordinal rating scale. The possible pairings in partial agreement then would logically receive scores somewhere between 0 and 1.

Capitalizing upon this phenomenon, Cicchetti and Sparrow[14] developed a general formula for obtaining a full set of linear weights for any size ordinal scale, as follows (where the symbol k refers to the number of ordinal categories on a given scale):

$$\text{Weights}_{\text{Linear}} = k–1, k–2, k–3, k–k$$
$$k–1, k–1, k–1\ k–1$$

For the 6-point aroma and bouquet wine ratings, the weights would become:

$(k–1)/(k–1) = 1$ for ratings in complete agreement,
$(k–2)/(k–1) = 4/5 = 0.80$ for ratings **one** scale category apart,
$(k–3)/(k–1) = 3/5 = 0.60$ for ratings **two** scale categories apart,

Table 11 How two wine tasters rate the acceptability of 100 wines

		Taster B	
Taster A	**Wine OK**	**Wine not OK**	**Totals/(Proportions)**
Wine OK	50	10	60 (0.60)
Wine not OK	5	35	40 (0.40)
Totals/ (Proportions)	55 (0.55)	45 (0.45)	100 (1.00)

$(k–4)/(k–1) = 2/5 = 0.40$ for ratings **three** scale categories apart,
$(k–5)/(k–1) = 1/5 = 0.20$ for ratings **four** scale categories apart and
$(k–6)/(k–1) = 0/5 = 0$ for ratings maximally apart on a 6-point ordinal scale.

Applying this weighting system to the data previously shown, and with PC determined in the same manner as for the 100 wines evaluated for whether they were considered appropriate for every day imbibing, we obtain the results shown in Table 13.

In the next two sections, we will apply the Winespider wine rating system, again to hypothetical data sets, in order to answer two interesting oenological questions: 1. What is the reliability or inter–taster agreement level when the same wines are evaluated currently and, say, 10 years hence? 2. Are there changes in the overall quality of the same wines as they change over the 10-year period of storage in an appropriately temperature and humidity controlled cellar?

How well do tasters agree, on overall wine quality, now and 10 years later?

Current inter-taster agreement on the overall quality of 10 wines

Assume the data present in Table 14 for the current assessment. When the same two tasters evaluate all the wines, model 2 of the ICC is applicable. The ANOVA that is appropriate is a two taster and 10 wines model.

Applying either the Chinese University of Hong Kong computer program (see http://department.obg.cuhk.edu.hk/researchsupport/IntraClass_correlation.asp) or the program developed by Cicchetti and Showalter,[17] produces the results shown in Table 15. The important point here is that the level of inter-taster agreement was 0.90, a result that is both statistically significant ($p < 0.001$), and clinically significant, or excellent by the criteria of Cicchetti and Sparrow;[14] and almost perfect, by the earlier criteria of Landis and Koch.[13]

Table 12 How two wine tasters rate the aroma and bouquet of 100 wines

							Rater 1
Rater 2	**O**	**A**	**P**	**G**	**VG**	**E**	**Totals**
1. Objectionable (O)	20	3	0	2	0	0	**25**
2. Acceptable (A)	2	10	3	5	0	0	**20**
3. Pleasant (P)	2	2	5	1	0	0	**10**
4. Good (G)	5	5	1	5	0	0	**16**
5. Very Good (VG)	1	1	2	1	10	0	**15**
6. Extraordinary (E)	2	1	1	0	0	10	**14**
Totals	**32**	**22**	**12**	**14**	**10**	**10**	**100**

The inter-taster agreement levels of the wines retasted 10 years later

Ten years later, the same wines receive, from the same two tasters, their overall evaluations (Table 16). Applying the same model of the ICC produces a value of 0.88, similar to the value of 0.90. It is also statistically significant at a probability level, p, of far less than 0.001. This value, as was true for the first tasting, is excellent by the criteria of Cicchetti and Sparrow,[14] and almost perfect[13] by Landis and Koch.[12]

The final hypothetical oenologic question, addressed in the next section, is whether there are statistically significant and/or clinically meaningful differences in the average overall ratings of the wines as they were tasted initially and then 10 years later?

Did the quality of the wines change over the 10 year period?

This straightforward question was addressed by testing statistically whether the average ratings of the 10 wines were statistically and clinically different over the 10 year period. The question was addressed by performing a paired t test, comparing the differences in the average ratings initially and at the 10 year mark. These differences were tested both for statistical and clinical meaningfulness (Table 17).

The results of the paired t test indicated no statistically significant differences between the two wine tastings in

Table 13 How well two raters agree on the aroma and bouquet of 100 wines

	Frequency Category Usage (%)	PO	PC	κ_w	Clinical significance
Taster 2					
1. Objectionable (O)	28.5	0.8456	0.5991	0.61	Good
2. Acceptable (A)	21.0	0.8238	0.6888	0.43	Fair
3. Pleasant (P)	11.0	0.8364	0.6913	0.47	Fair
4. Good (G)	15.0	0.7067	0.6491	0.16	Poor
5. Very good (VG)	12.5	0.9040	0.5416	0.79	Excellent
6. Extraordinary (E)	12.0	0.8583	0.3893	0.77	Excellent
Overall	**100**	**0.8280**	**0.6032**	**0.57**	**Fair**

Notes: These data indicate the following: 1. The level of agreement, averaged over the six ordinal categories of aroma and bouquet evaluations is acceptable, or Fair, at a chance-corrected, or κ_w value of 0.57. 2. On a category by category basis, there is a wide array of inter–taster reliability levels. These range from Poor (0.16) for a rating of Good aroma and bouquet; to Fair, for the categories Acceptable (0.43) and Pleasant (0.47); to Good for Objectionable (0.61); to Excellent for both Very good (0.79); and Extraordinary (0.77) levels of aroma and bouquet. This variation is not surprising, since, in general, across many types of ratings, the extreme categories tend to be more reliably rated than categories tending toward the middle of a given rating scale. This is perhaps due to the enhanced salience of extreme categories of classification on any given ordinal scale.

terms of average taster scores, at or beyond the conventional $p = 0.05$ level of statistical significance. The value of t was only 0.33, and a value of 2.26 was required to reach statistical significance at the conventional $p = 0.05$ level. This means the average differences in the tasters' wine ratings initially and 10 years later are best interpreted as chance findings rather than either a statistical or clinically meaningful level.

These findings are also consistent with the results of a careful inspection of the data, where it can be seen that the only wine that changed appreciably was Wine A, whose average score increased from 72.5 (Acceptable, Fair range) to 82.5 (in the Good range). The overall result is also very consistent with the average of the average ratings being 83 at the first tasting and about 85 10 years later.

Finally, the results are consistent with the well known oenologic fact that except in rare instances, wines do not, on average, get appreciably better in quality, over time.

Table 14 Overall initial ratings of 10 wines by two wine tasters

Wine	Taster A	Taster B
A	70	75
B	85	80
C	67	72
D	84	85
E	92	90
F	88	85
G	92	95
H	86	87
I	79	77
J	84	88

Summary, conclusions, and implications for future research

In this report, we provided criteria for selecting, among many that were available, those wine rating scales that appear useful to wine producers, wine consumers, and oenologic researchers. We discussed the anatomy or internal structure of each of these rating scales, followed by their application to answer a series of oenologic questions that would also be of interest to producers, consumers, and wine researchers.

As a final caveat, we need to understand that wine scores cannot be taken as absolute and inflexible. Just as one can expect there to be test-retest variability in the application of clinical measurements in other areas of scientific inquiry, such as IQ testing, performed by the most skilled of cognitive specialists, we should expect, on average, a consistent and measurable amount of inter-taster variability in the rating of the same wine by two or more highly experienced independent tasters. This has especial relevance when a single source, say a professional wine taster, offers a score that straddles two levels of suggested wine quality. Thus, a score of, say, 89 (Very good) could very easily become a score of 90 or one that defines the same wine as Excellent. Unfortunately, to date this critical problem appears to not have been addressed, as far as we are aware. It is, in our judgment, an issue that needs to be carefully considered and measured in future oenologic research.

One of the little studied factors that may contribute considerably to both intra-taster and inter-taster variability is an issue that the authors described in a very recent article,[18] namely the oenologic observation that although

Table 15 Analysis of variance (ANOVA) of overall wine ratings by 2 hypothetical tasters of 10 wines at first tasting

Source	Degrees of freedom (df)	Sums of squares (SS)	Mean square (MS)	F Ratio
Between tasters (T)	1	2.4531	2.4531	0.39 (NS)
Between wines (W)	9	1055.4530	117.2726	18.5015
T × W	9	57.0469	6.3385	
Totals	19	1114.9530		

$$\text{ICC} = \frac{\text{MSS}-\text{MSE}}{\text{MSS} + (\text{MSE})(t-1) + t(\text{MST}-\text{MSE})/N} = 0.90.$$

Notes: The level of statistical significance of this ICC value is determined by the size of the F ratio for between wines. The value of 18.50 is statistically significant at far beyond the probability of $p < 0.001$. Also, the ICC is excellent by the clinical criteria of Cicchetti and Sparrow,[14] and almost perfect by the earlier strength of agreement criteria of Landis and Koch.[13]
Abbreviations: MSS, mean square between wines; MSE, mean square of TXW; MST, mean square between tasters; t, the number of tasters; N, the number of wines.

the recognition threshold level for sugar is between 0.5% and 2.5%, the average taster recognition is only 1%. Consequently, there may be considerable variability in the perception of sugar between any two tasters. What one taster may detect as sweet may seem dry to another, despite the fact that they taste the same wine!

It is also important to note that for the sake of simplicity, we have focused here on the situation in which two tasters evaluate a number of wines. For detailed discussions of how these same kappa-type statistics are utilized when there are multiple wine judges or tasters, the interested reader is referred to the work of Cicchetti,[19–21] using 11 wine tasters; as well as the very recent work of Hodgson,[22] in which panels of four judges were used to evaluate whether a wide range of wines qualified for a Gold, Silver, or Bronze medal, or no medal at all.

Finally, we should like to complete this wine essay by referring to three incisive comments, in the form of queries, that an anonymous reviewer raised for their clear heuristic value for further research pertaining to some of the basic issues we have raised: 1. How might the oenologic researcher weight the relative importance of the 16 wine attributes in the Winespider evaluation system? 2. Is there hope, eventually, for a "gold standard" approach to assessing the validity or accuracy of blind wine tasting? 3. Is it possible to achieve ratio scaling methods in the further construction of wine rating scales of the future?

We are grateful to the anonymous reviewer for raising these important questions that will require considerable thought in the design of future oenologic investigations, but questions that we are convinced can lead the field farther forward in the quest for an ever more secure scientific foundation.

Disclosure

The authors report no conflicts of interest in this work.

Table 16 Overall quality ratings of 10 wines by two raters 10 years later

Wine	Taster A	Taster B
A	80	85
B	87	89
C	72	75
D	78	80
E	90	90
F	90	85
G	90	92
H	87	87
I	82	83
J	85	84

Table 17 Average ratings of 10 wines by two raters, initially and 10 years later

Wine	Average initial rating	Average rating 10 years later
A	72.5	82.5
B	82.5	88.0
C	69.5	73.5
D	84.5	79.0
E	91.0	90.0
F	86.5	87.5
G	93.5	91.0
H	86.5	87.0
I	78.0	82.5
J	86.0	84.5
Averages	**83.05**	**84.55**

References

1. Amerine MA, Roessler EB. *Wines: Their sensory evaluation.* New York: WH Freeman; 1983.
2. Kaufman, AS. Do low levels of lead produce IQ loss in children? A careful examination of the literature. *Arch Clin Neuropsychol*. 2001;16:303–341.
3. Cicchetti DV, Showalter D, Tyrer P. The effect of number of rating scale categories upon levels of interrater reliability: A Monte Carlo investigation. *Appl Psychol Meas*. 1985;9:31–36.
4. Preston CC, Colman AM. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*. 2000;104:1–15.
5. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;23:37–46.
6. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull*. 1969;72:323–327.
7. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968; 70:213–220.
8. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Reports.* 1966;19:3–11.
9. Bartko JJ. Corrective note to: "the intraclass correlation coefficient as a measure of reliability." *Psychol Reports*. 1974;34:418.
10. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*. 1975;31:651–659.
11. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33:613–619.
12. Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: Wiley: 1981.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
14. Cicchetti DV, Sparrow SS. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *Am J Ment Defic.* 1981;86:127–137.
15. Cicchetti DV. Heavens R.RATCAT (Rater Agreement/Categorical Data). *Am Stat.* 1979;33:91.
16. Heavens RH Jr, Cicchetti DV. A computer program for calculating rater agreement and bias statistics using contingency table input. *Proc Amer Stat Assoc (Stat Computing Sec)*. 1978;21:366–370.
17. Cicchetti DV, Showalter D. A computer program for determining the reliability of dimensionally scaled data when the numbers and specific sets of examiners may vary at each assessment. *Educ Psychol Meas*. 1988;48:717–720.
18. Cicchetti A, Cicchetti D. The balancing act in consistent wine tasting and wine appreciation: A tale told by two brothers. Part 1: Consistency in wine tasting and appreciation: a personal-experiential perspective. *J Wine Res*. 2009;19:115–121.
19. Cicchetti DV. Who won the 1976 blind tasting of French bordeaux and American cabernets? Parametrics to the rescue. *J Wine Res*. 2004;15:211–220.
20. Cicchetti DV. The Paris wine tastings visited once more. *J Wine Econ*. 2006;1:125–140.
21. Cicchetti DV. Assessing the reliability of blind wine tasting. *J Wine Econ*. 2007;2:196–202.
22. Hodgson RT. An examination of judge reliability at a major US wine competition. *J Wine Econ*. 2009;3:10–113.