On the Reliability and Accuracy of Wine Tasting: Designing the Experiment; Selecting the Tasters; and Assessing and Interpreting the Resulting Data

Dom Cicchetti, Ph.D.
Department of Biometry
Yale School of Medicine
New Haven, CT 06471

Abstract

This report has as its primary goal a discussion of the ratings of wine within a more comprehensive framework of an appropriate wine tasting research paradigm, - one that takes into account the following: the characteristics of the sample of judges; the type of a priori screen required to eliminate unreliable inconsistent wine tasters before the tastings/competition begins; a studied selection of the wine qualities that the tasters will be asked to apply; the selection of appropriate wine glasses; the selection of well-defined measurement scales; and the application of standard and state-of-the-art biostatistical procedures, as required, to analyze and interpret the results of the tastings.

Research Objective

The purpose of this research is to discuss open or blind wine tasting in a comprehensive context of: Designing an appropriate experiment; assessing the reliability and accuracy of the tasting; and interpreting the meaning of the results.

Designing an Appropriate Experiment

An important consideration here is whether to blind the wine tasters or not. Some insist, with plausible arguments, that blinding is necessary to eliminate potentially biasing cues, such as knowing the varietal(s); who the vintners are; where the wine comes from (its so-called terroir); the cost of the wine; etc.

That such factors do indeed influence the results of a wine tasting was clearly demonstrated in a well-designed study by Prilaid (2007). When raters knew the origin of the wine, it influenced the rating of the quality of the wine; when the location was not made available, the quality of the wine was not so influenced.

There are other proponents, who are of the opposing opinion. They aver, again with convincing arguments, that there are critically important qualities of a wine, such as color intensity and aroma that need to be made available to the taster in order to render a valid rating. Be that as it may, the wine tasting experiment discussed here is applicable to both blind and open tastings. As noted earlier, (Cicchetti, 2004b), the factors that are to be considered in a given wine tasting experiment, include:
1. **How the sample of wine tasters is chosen**

2. **The type of a priori screen required to eliminate unreliable (inconsistent) wine tasters**
   before the competition begins.

3. A careful selection of **the specific wine qualities that the tasters are required to utilize**
   (with specific criteria provided to the wine tasters). As one excellent example, Cliff & King
   (1996) used a modified wine rating scale that had previously been developed by
   Amerine & Roessler (1983). As a specific application, they provided the following criteria
   to each wine judge, for evaluating *aroma and bouquet*:

   6= extraordinary. Unmistakable characteristic aroma of grape variety or wine type.
   Outstanding and complex 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= poor. With slight off odors.
   0= objectionable. Objectionable. with off odors.

4. **Choosing appropriate wine glasses.** While Mr. Riedel has been conducting informal
   “experiments” to demonstrate how much better a wine tastes when it is drunk from an
   appropriate crystal wine glass- (the author participated in one of them)-a more formal
   research study was designed by the outstanding wine researcher Margaret Cliff (2001) that
   confirms empirically Riedel’s more informal experiential findings. As noted in Cicchetti
   (2004), p.224: “In her classic research design, …. Cliff (2001), was able to demonstrate
   strong linear relationships between such wine characteristics as total intensity and colour,
   and the physical dimensions of the glasses: opening diameter, maximum diameter (cuppa),
   height and volume.

   As Cliff correctly concludes, her research supports the use of a specific type of glass to
   evaluate a specific varietal of wine.”

5. **Choosing well-defined wine measurement scales.** As noted in Cicchetti & Cicchetti (2009),
   there are a number of extant wine rating scales. Some of the more important ones include:
   My Wine Rating scale; the aforementioned Amerine & Roessler (1983) wine rating system;
   the red wine buzz rating system; the Jancis Robinson wine rating scale; and others
   developed by the Wine Spectator; Robert Parker; Steven Tanzer; the Nick Chebnikowski
   Winespider evaluation system, and, finally, there is the important issue of:

6. **Selecting and applying standard and state-of-the-art biostatistical procedures to analyze
   and interpret the enological results.** In the next section, I shall discuss research on
   assessments of the reliability of wine tasting among the experts (here read Jancis Robison
   and Robert Parker) for both the 2004 Bordeaux wines (Cicchetti & Cicchetti, *in press*,
Reliability of Ratings of the 2004 and 2009 Bordeaux Wines

In the “informal literature”, two titans in the wine world are Jancis Robinson (JR) from England, and America’s Robert Parker (RP). RP has been depicted as one who is enamored with lush, super-ripe so-called “fruit-bombs” while JR is touted as one who appreciates wines that are much more subtle; much “less” fruit-up-front”, while also evidencing overall balance. This characterization of these proclaimed enological experts may help to explain the profound disagreement between the two in the rating of close to 2004 Bordeaux wines (184, to be precise), in the 2004 tastings.

Using standard and state-of-the-art biostatistical techniques, Cicchetti & Cicchetti (in press) showed the following: Applying the weighted kappa statistic developed by Cohen (1968), using the corrected standard error theoretically justified by Fleiss, Cohen, & Everitt (1969); and empirically verified by Cicchetti & Fleiss (1977), the following results were obtained (i.e., Cicchetti & Cicchetti, in press, Journal of Wine Research):
The overall level of agreement was exceedingly low, barely exceeding chance-expectations. Analyzing these results further, using the test of rater bias introduced by McNemar (1947) and generalized by Bowker (1948):

There was complete agreement on only 39% of the 184 wines; Of the 113 disagreed upon wines, RP gave a higher rating than did JR on 110 (97%) of them. RP scored: 51 wines as Excellent that JR rated as Average; 38 as Above Average that JR scored as Average; 19 as Excellent that JR rated as merely Average; and 2 wines as Above Average that JR rated as Below Average.

The results of the 2009 tastings were even more dismal, as reported recently by Cicchetti & Cicchetti (2012). Utilizing the same biostatistical approaches applied to the aforementioned 2004 data, produced the following results: Of the 237 Bordeaux they evaluated, RP and JR were in agreement on only 65 or 27% of them, thereby resulting in disagreement on 73% or nearly 3/4 of the remaining wines. Finally, of the 172 disagreed upon scorings, RP gave a higher rating than JR on every one of them, thereby producing the maximum amount of rater bias that was possible.

Some recent follow-up data suggests the possibility that groups of tasters share similar likes and dislikes of wines. For example, while the agreement level between JR and RP on both 2004 and 2009 Bordeaux wines is almost embarrassingly low, the agreement level between RP and James Suckling (JS) of the Wine Spectator staff, on the 2004 Bordeaux was respectable at .52, or 76% (Cicchetti, unpublished).

What is of interest here, at least from an experiential perspective, is the known similarity in ratings that are often made of the same wine by RP, JR, JS, and Steven Tanzer. All this said, future empirical research will be required to determine how broad are the implications of such a finding.

But what of the broader implications of the concept of the reliability of blind wine tasting? Is there any evidence to suggest that there may exist among us some wine tasters who actually evidence acceptable levels of inter-judge agreement? And ones that are also reproducible results?

This question was posed by the author several years ago:

In re-analyses of famed 1976 Paris and California taste-off of Bordeaux/California Cabernets; and French and California Chardonnays (Cicchetti, 2004a; and 2006), it was shown for the first time that the overall results, among a distinguished group of wine tasters, was, in fact, far less than distinguished:

The overall results among all tasters revealed a mere chance-corrected agreement level of .36 (or 68%) for the Chardonnay scorings; and an even more dismaying level of .22 (or 61%) for ratings of the Bordeaux/Cabernets.
Realizing that this was an omnibus result, the question was asked: Is it possible that within the group of 11 tasters, there might be two subsets, one with respectable levels of reliability (Agreement of at least 70%); and, even more importantly, would the results be reproducible? The answer to this second question was built into the Paris-design of the research, since the very same set of tasters imbibed both the Chardonnay and Cabernet wines. Using a new technique (Cicchetti, 2004a), demonstrated that:

There were indeed two subsets of wine tasters:
- 5 of the 11 tasters had a chance-corrected Agreement level of .73 (or 86.5%), for the Chardonnay wines; and a comparable .69 (or 84.5%) for the Bordeaux/Cabernets.
- The remaining 6 tasters had unacceptable levels of .36/68%, for the Bordeaux/Cabernet wines; and .30/65% for the Chardonnays.

The final issue that will be discussed here is how to determine the accuracy or validity of wine tastings.

Determining the Accuracy of Wine Tasting

The concepts of inter-rater reliability and accuracy of judgments (as in the case of wine scorings), are -as we learned in biostatistics 101-intrinsically related, or intertwined. As we were taught, if a phenomenon is measured accurately/validly, it is per force also a reliable/consistent one. The reverse, however, does NOT hold. Thus, my weight on my scale might be consistent/reliable, albeit incorrect/invalid/inaccurate, when compared to, say, the presumably gold-standard weight provided by the scale in my Doctor’s office.

We have just seen that some tasters do, indeed, provide reliable wine ratings; and, in fact, it is not a leap in logic to believe that it is possible for experienced, reliable, wine tasters to teach their inexperienced enological brethren to also score wines reliably.

But how about the accuracy or validity of wine ratings? How can this phenomenon be assessed appropriately? Cicchetti & Cicchetti (2009) assessed the utility of major extant wine rating scales for producers, consumers, and enological researchers. The scale that proved to be the most useful and comprehensive, and therefore, most accurate, was the Winespider Evaluation System developed by an artist of repute, named Nick Chebnikowski (Chebnikowskiwww.winespider.com).

The scale has the following characteristics:
- It contains 16 wine attributes: color, viscosity, brilliance, depth, aroma, faults, varietal, intensity, complexity, concentration, fruit, length, aftertaste, balance, tannins, and acid.
- Each attribute is measured on a 1-10 point ordinal scale; so that the total possible score for a “perfect” wine is 16 X 10= 160.

Since the 16 ratings form a spider web pattern, it was defined, eno-eponymously, as the Winespider Evaluation system. and, finally,

This is the only extant wine rating scale that allows for tracking any changes in the 16 wine attributes over time, as the wine ages. In conclusion, validity or accuracy of ratings, in the context of wine tasting, has to do with the extent to which the scale depicts all those characteristics of wines that are relevant. To date, the Winespider Evaluation System comes closest to meeting this challenging criterion.

Summary

This report sought to discuss the judged ratings of wine in a more comprehensive framework of an appropriate wine tasting research paradigm, taking into account the following: the
characteristics of the sample of judges; the type of *a priori* screen needed to eliminate unreliable/inconsistent wine tasters before the tastings/competition begins; a studied selection of the wine qualities that the tasters will be asked to apply; the selection of appropriate wine glasses; the selection of well-defined measurement scales; and the application of standard and state-of-the-art biostatistical procedures, as required, to analyze and interpret the results of the tastings.

It was next shown that while some research indicates that wine experts display dismal, quite meaningless levels of agreement, there is also evidence to indicate that other judges have demonstrated quite adequate levels of agreement; and ones that have replicated successfully. It is also possible and even likely, that groups of tasters share similar likes and dislikes of wines. For example, while the agreement level between JR and RP on both 2004 and 2009 Bordeaux wines is almost embarrassingly low, the agreement level between RP and James Suckling (JS) of the Wine Spectator wine staff, on the 2004 Bodeaux was respectable at .52, or 76%. Future research will be required to determine how broad are the implications of such a finding.

References


