Journal of Business and Economics, ISSN 2155-7950, USA
August 2014, Volume 5, No. 8, pp. 1199-1204
DOI: 10.15341/jbe(2155-7950)/08.05.2014/001

http://www.academicstar.us

Categorizing A Wine Rating Scale: 2, 3, 4, or More: Is There One We Should Go For?

Dom Cicchetti[1], *Arnie Cicchetti*[2]

(1. Department of Biometry, School of Medicine, Yale University, New Haven, CT 06520, USA; 2. San Anselmo, CA 94960, USA)

Abstract: The purpose of this research is to provide criteria for selecting wine scales that are not only useful for researchers, as well as the wine trade, but also for the everyday consumer. In discussing the relative merits and flaws of each of the wine scales, we have stressed that because of the relative nature of this endeavor, we must remain flexible. There are no absolutes. Variability not only exists in a wine, a living organism, but also among the most experienced tasters. We present the findings of an earlier Monte Carlo study of the relationship between the number of categories or scale points and a given scale's level of inter-rater agreement. It has been demonstrated empirically that reliability increases dramatically from 2 to 3 scale points, and continues to increase linearly up to 7 scale points, where a leveling off occurs, such that no appreciable levels of reliability occur with increases in scale points, even when they reach as many as 100 (Cicchetti, Showalter, & Tyrer, 1985). Applying these results to the selection of wine rating scales, we would eschew the 3 category whimsical 3 Stooges scale, with its categories defined by fantasized Moe, Larry, and Curly wine descriptors. As given in earlier research, we offer a rationale for choosing the Winespider Evaluation System, developed by the Australian artist, Nick Chlebnikowski, as our "gold standard" for the most reliable and valid extant wine rating scale (www.winespider.com) Cicchetti & Cicchetti (2009).

Key word: wine rating scales

JEL code: C020

1. Introduction

Whether it be a wine judge, winery principal, teacher, wine educator or consumer, it appears that everyone has a special interest in how wine not only tastes, but how well it will be appreciated in years to come. Many not only rely on their palates, but turn to wine specialists who have created wine scales applying a numerical value to each bottle of tasted wine.

In this report, the authors provide criteria for selecting among the many, those wine scales that appear to be useful to producers, collectors, educators, and everyday consumers. We discuss the anatomy or internal structure of each of these scales, offering examples in application that might be of interest to the industry.

Of course, we are dealing with living organisms and as such we cannot become inflexible to the wine's

Domenic Cicchetti, Ph.D., Professor, Yale University; research areas/interests: reliability and accuracy of human judgment, autism, general diagnostic issues, test development, wine research. E-mail: dom.cicchetti@yale.edu.

Arnie Cicchetti, B.A., Consulting Specialist of Wine (CSW), and past Director of National Accounts, Wine Warehouse, San Anselmo, CA 94960. E-mail: acicchetti@earthlink.net.

propensity for growth. So too, we must allow for the variability between tasters, and the goal that each of these scales is trying to achieve. Just as there can be expectations of variability in a test, re-test scenario in a clinical laboratory, one must accept the same phenomenon in the tasting of wine. To that note, what is of more importance to the authors is consistency in these inter-variable ratings. Without consistency, a numerical value is useless. For example, the difference in scoring of one point (say 89-90) can on some scales prove to be the difference between a wine scored "Very Good (89)", and a wine scored "Excellent (90)". Thus, with one point the wine straddles two separate levels of suggested wine quality. To date, however, we have not found in our research, this critical problem being addressed.

2. Objectives

The objectives of this paper are: (1) to provide an overview of the major wine rating scales; (2) to describe their structure or anatomy and; (3) to give hypothetical examples of how their reliability would be assessed, and the information thereby obtained; and (4) we offer our "Gold Standard" for the available wine rating scales.

3. The Anatomy of Wine Rating Scales

The anatomy of the wine scales is divided into the following: Nominal/Categorical; Ordinal (Rank-ordered Scales); and Interval/Interval-Made-Ordinal-Scales.

3.1 Nominal/Categorical Rating Scales

This simplest type of scale can be classified, broadly, under one of two types of categories: those that are biologically determined, and therefore have a predetermined or fixed number of categories, such as gender defined as female or male, or the stages of a disease, defined by specific and reproducible criteria; and the much more frequent type of categorical scale, wherein the number of categories is indeterminate, such as the sensory terms defined by Alonso et al. (2010).

A somewhat whimsical wine rating scale that is of dubious scientific worthiness and of anonymous authorship is categorized as the Three Stooges Wine rating scale. The three scale descriptors that define this wine rating scale are:

MOE = A wine that is crude, harsh tasting, tannic, acidic, and bops your tongue with a closed fist!

LARRY= A wine that is easy going, inoffensive, soft, and trying hard not to grate

CURLY = A wine of great character and distinction.

A much more scientifically serious example used is Odor Intensity as defined carefully, on the basis of well-defined sensory perception criteria by Alonso et al. (2010). In this rating scale the sensory terms such as herbaceous, lactic, tree fruit smells, etc are used to determine quality. There is no numerical value attributed to the wine. In this rating one relies solely on sensory perception to determine the value of the wine.

3.2 Ordinal Ranking

One such wine rating scale relies on a more personal ranking (e.g., My Wine Rating Scale, (Tim, 2005)). This anonymously designed scale consists of 7 rankings ranging from < 4 (Undrinkable) to 10 (Excellent). The values are rather non-definable, or non-criterion specific, such as a rating of 5 (Pretty Bad), or a rating of 7 (Quaffable), and thus limiting in any public discussion. This scale seems to follow the premise that the less said about wine, the better. However, Tim may just desire to keep any knowledge that he possesses about wine to himself. This may be an admirable trait but makes it quite limiting for serious scientific discussion.

A second, and in our judgment, scientifically defensible wine rating scale, is referred to as the Red Wine-Buzz scale, with a cyberspace address of redwinebuzz.com. This scale rates 9 wine characteristics on 5 point ordinal scales; (and overall quality, on a Six Category Ordinal Scale):

(1) Color
(2) Nose
(3) Palate
(4) Finish
(5) Tannins
(6) Acidity
(7) Alcohol
(8) Aging Potential and
(9) Food Friendliness

The Red wine buzz uses well-defined Scale Descriptors, such as the criteria for rating *palate* as the following:

1 = Very Vague/Simple Flavors
2 = Straightforward Flavors
3 = Medium complexity of Flavors
4 = Complex Flavors
5 = Very Complex & Persistent Flavors

A third example of a commendable ordinal category wine rating scale is one published in 1983 by Amerine & Roessler. Ratings, with specific criteria, are designated as shown here:

(1) Appearance & Color (0-2)
(2) Aroma & Bouquet (0-6)
(3) Total Acidity (0-1)
(4) Balance (0-2)
(5) Body (0-1)
(6) Flavor (0-3)
(7) Finish (0-2)

The following is an example of the Criteria for the Ordinal Rating of Aroma & Bouquet:

1 = Objectionable, with or without off-odors
2 = Acceptable without perceptible Aroma or Bouquet
3 = Pleasant with slight Aroma or Bouquet
4 = Good with characteristic Aroma, and distinguishable Bouquet
5 = Very good with characteristic aromas and complex Bouquet
6 = Extraordinary unmistakable characteristic Aroma of a grape varietal or wine type. Outstanding & complex Bouquet. Exceptional balance of Aroma Bouquet.

3.3 Interval and Interval-Made-Ordinal-Scales

Two of the more prominent scales/ratings used in this ranking are:

Overall Interval Scale Rating/Made Ordinal (Parker/Wine Spectator): The Wine Spectator, scale ranges between 50-100 points. Although the Wine Spectator considers 60-69 point wines as drinkable, they do not recommend any wine scored below 70 points. Parker, on the other hand, does not *rate* any wine below 70 points.

Thus these two wine rating scales are similar in ratings from 70 points up to 100 points. Their terminology is different, however.

When defining the wines from 80-89 points, The Wine Spectator lists this category as Good: Solid; Well-Made. Mr. Parker groups these categorized wines as Above Average Quality. To the consumer, however, these terminologies are the same. As previously stated, both publications have a cut-off point of 89 for Good/Above Average wine. One point more in the reviewers' mind for either publication results in an entirely different classification: Outstanding/Superior Quality.

This has been seen by many a winemaker/winery owner as unfair. Couple that with the consumer's desire to want to drink the "best" wine available, placing wines with a score of less than 90 points on many consumers' "non-preferred" list, can create an obvious problem. To date this critical problem appears to not have been addressed by any publication doing Interval and Interval-Made-Ordinal-Scales.

Another publication listed below, from Jancis Robinson (JR) uses a different ordinal scale. This range is: 12-20 points.

12-13 = Below Average Quality

14-15 = Average Quality

16-17 = Above Average Quality

18-20 = Superior Quality

As stated in Cicchetti & Cicchetti (2009), the JR scale can be equated to other 100 point scales by simply using multiples of 5 for each JR rating.

As noted earlier (Cicchetti & Cicchetti, 2009), the most comprehensive Ordinal-Interval Scale is The Wine Spider Evaluation System (Chebnikowski, www.winespider.com). This evaluation system contains: 16 wine attributes: color, viscosity, brilliance, depth, aroma, faults, varietal, intensity, complexity, concentration, fruit, length, aftertaste, balance, tannins, and acids.

(1) Each attribute is measured on a 1-10 scale

(2) Total score = 16 x 10 = 160 possible points

(3) The 16 ratings form a spider web pattern

(4) This is the only extant scale that tracks changes in the wine as it ages.

The next part of this research is devoted to a fundamental research question: As the number of scales categories increases, does inter-rater agreement increase, decrease, or stay the same. In order to answer this question we must turn to Cicchetti, Showalter, & Tyrer (1985). The authors designed a Monte Carlo Investigation on: The effect the number of Rating Scale Categories had on levels of Inter-rater Agreement/Reliability. Random pairs of Raters were simulated, using ratings varying in:

(1) The number of scale points: 2-10; 15, 30, 50; 100

(2) The percentage of absolute agreement: 50%; 60%; 70%

(3) The proportion of cases when one Rater gave higher ratings than the other when they disagreed: (50-50; 60-40; 70-30; 90-10)

(4) The sample size for each scale was 200

(5) There were 10,000 computer simulations for each experimental condition.

The next step was to use a conversion formula for the Reliability Statistic of choice, the Intra-Class Correlation Coefficient (ICC).

Robinson (1957), in what might be referred to as an oldie-but-goodie scientific publication, showed that the

(ICC) for two raters can be converted into an agreement statistic by the following simple conversion formula: Percentage agreement (% A) = (ICC+1)/2.

The advantage of this straightforward conversion is that "A" is easier to interpret than ICC. Applying this conversion, the following results occurred (Cicchetti et al., 1985): Reliability & Number of Scale Points: A = 60%, Bias = 60/40

K	A(%)	Clinical Significance
2	60.5	Poor
3	70	Fair
4	73.5	Fair
5	75	Fair
6	76	Fair
7	76.5	Fair
8	77	Fair
9	77.5	Fair
10	77.5	Fair
15	78.5	Fair
30	79	Fair
50	79.5	Fair
100	79.5	Fair

Reliability & Number of Scale Points: A = 50%; Bias = 60/40

K	A(%)	Clinical Significance
2	50.5	Poor
3	62.5	Poor
4	66.5	Poor
5	69	Poor
6	70	Fair
7	71	Fair
8	71.5	Fair
9	72	Fair
10	72.5	Fair
15	73	Fair
30	74	Fair
50	74.5	Fair
100	75	Fair

It should be noted that the same pattern of results held for the remaining levels of simulated rater agreement levels of (70%; 80%; 90%); and the remaining levels of simulated rater bias (70%/30%; 80%/20%; and 90%/10%).

4. Discussion and Conclusion

Reliability increases dramatically from 2 to 3 scale points. It then increases upwards to **six or** seven

categories, leveling off so that a 100 category scale is no more meaningfully reliable than a seven category scale. These results were replicated successfully in a later investigation by Preston & Colman (2000).

(1) In selecting a wine rating scale, the categories need to be well defined, non-overlapping, and the raters must be trained to produce reliable wine ratings.

(2) To achieve validity and accuracy the scale needs to have wide coverage (please refer to The Wine Spider Evaluation System, as shown in Cicchetti & Cicchetti, 2009). This wine rating scale is our "Gold Standard".

There is also an important caveat: Any Wine Rating Scale (including The Three Stooges product of whimsy) can be made reliable through proper training of the Raters. However, because of its poor coverage in dealing with enological detailing, such a wine rating scale would not be very accurate or valid.

In summary, the purpose of this paper is to provide criteria for selecting wine scales that are not only useful to researchers, and the wine trade, but also to the everyday consumer. In discussing the relative merits and flaws of each of the wine scales, we have stressed that because of the relative nature of this endeavor, we must remain flexible. There are no absolutes. Variability not only exists in a wine, a living organism, but, too, in the most experienced tasters.

Next, the categories of a wine scale must be well defined, and non- overlapping. The scale also needs to have wide coverage.

Finally, the reader interested in a detailed description of how to assess the reliability of rating scales, as they are applied to the evaluation of the perceived quality of wine, should again consult Cicchetti & Cicchetti (2009).

References:

Alonso I. E. (2010). "A method of development for sensory quality control of products with certified quality labels: A case study on wine", in: *Meeting of Sensometrics*, Rotterdam, The Netherlands.

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

Cicchetti D. and Cicchetti A. (2008). "The balancing act in consistent wine tasting and wine appreciation: A tale told by two brothers—Part II: Consistency in wine tasting and appreciation: An empirical-objective perspective", *Journal of Wine Research*, Vol. 19, pp. 185-191.

Cicchetti D. V. and Cicchetti A. F. (2009). "Wine rating scales: Assessing their utility for producers, consumers, and oenologic researchers", *International Journal of Wine Research*, Vol. 1, pp. 73-83.

Cicchetti D. V., Showalter D. and Tyrer P. (1985). "The effect of number of rating scale categories upon levels of interrater reliability", *Applied Psychological Measurement*, Vol. 9, pp. 31-36.

Preston C. C. and Colman A. M. (2000). "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences", *Acta Psychologica*, Vol. 104, pp. 1-15.

Robinson W. S. (1957). "The statistical measurement of agreement", *American Sociological Review*, Vol. 22, pp. 17-25.

Tim (2005). "My wine rating scale", unpublished tasting notes.