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

Dom Cicchetti, Ph.D., Yale University Department of Biometry

Arnie Cicchetti, Certified Specialist of Wine (CSW)

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 exits 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 increases 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).

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 can not become inflexible to the wine’s 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.

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 finally, we offer our ‘Gold Standard” for the available wine rating scales.

The anatomy of the wine scales is divided into the following:

1) Nominal/ Categorical
2) Ordinal (Ranked Scales)
3) Interval/Interval-Made-Ordinal-Scales
Nominal/Categorical Rating Scale
The example used is Odor Intensity (Alonso, et al., 2010) In this rating scale the authors use sensory terms such as herbaceous, lactic, tree fruit smells, etc 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.

Ordinal Ranking
This system relies on a more personal ranking (My Wine Rating Scale, (Tim, 2005). It consists of 7 rankings from <4 (Undrinkable) to 10 (excellent). The values are rather non-definable- 5 (Pretty Bad), 7 (Quaffable), and thus limiting in any public discussion! This scale takes 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 limiting for discussion!

Interval and Interval-Made-Ordinal-Scales
Two of the more prominent scales/ratings used in this ranking are:

Overall Interval Scale Rating (Parker /Wine Spectator): 70-100 points, and Made Ordinal: The Wine Spectator, 50-100 points. Although the Wine Spectator ranks 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 giant publications 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. 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 (2008), the JR scale can be equated to other 100 point scales by simply using multiples of 5 for each JR rating.

The most comprehensive Ordinal-Interval Scale is The Wine Spider Evaluation System (Chebnikowski www.winespider.com). This evaluation system contains:

1) It contains 16 wine attributes: color, viscosity, brilliance, depth, aroma, faults, varietal, intensity, complexity, concentration, fruit, length, aftertaste, balance, tannins, and acids.
2) Each attribute is measured on a 1-10 scale
3) Total score = 16 x 10 =160 possible points.
4) The 16 ratings form a spider web pattern
5) This is the only extant scale that tracks changes in the wine as it ages.
We next look at Categorical Scales:

These are:
1) Number of categories biologically determined: example, gender-male and female, or disease stages (0,1,2,3,4)
2) Number of categories indeterminate: (e.g., pain, happiness, clinical depression).

Wine ratings fit into Category #2, Indeterminate Classification.

Other Wine Rating Systems

1) The Three Stooges Wine Rating System

(This enters the category of whimsical/ridiculous).

Scale Descriptors:

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.

The Redwinebuzzscale at redwinebuzz.com.

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

Redwinebuzz Scale Descriptors (e.g The Criteria for rating palate)

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


Ratings, with specific criteria:

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 #2 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 or Aroma Bouquet.

The next part of this research is devoted to a fundamental research question: As the number of scales categories increases, does the 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. 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: Intra-Class Correlation Coefficient (ICC) Robinson (1957) showed that the (ICC) for two raters can be converted into an agreement statistic by the following simple formula:
Percentage agreement (A%)=(ICC+1)/2
Advantage in this: “A” is easier to interpret than ICC.
With this conversion the following results occurred
(Cicchetti, et al., 1985):

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

<table>
<thead>
<tr>
<th>K</th>
<th>A(%)</th>
<th>Clinical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>60.5</td>
<td>Poor</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>Fair</td>
</tr>
<tr>
<td>4</td>
<td>73.5</td>
<td>Fair</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>Fair</td>
</tr>
<tr>
<td>7</td>
<td>76</td>
<td>Fair</td>
</tr>
<tr>
<td>8</td>
<td>76.5</td>
<td>Fair</td>
</tr>
<tr>
<td>9</td>
<td>77.5</td>
<td>Fair</td>
</tr>
<tr>
<td>10</td>
<td>77.5</td>
<td>Fair</td>
</tr>
<tr>
<td>15</td>
<td>78.5</td>
<td>Fair</td>
</tr>
<tr>
<td>30</td>
<td>79</td>
<td>Fair</td>
</tr>
<tr>
<td>50</td>
<td>79.5</td>
<td>Fair</td>
</tr>
<tr>
<td>100</td>
<td>79.5</td>
<td>Fair</td>
</tr>
</tbody>
</table>
Reliability & Number of Scale Points: A=50%; Bias=60/40

<table>
<thead>
<tr>
<th>K</th>
<th>A(%)</th>
<th>Clinical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>50.5</td>
<td>Poor</td>
</tr>
<tr>
<td>3</td>
<td>62.5</td>
<td>Poor</td>
</tr>
<tr>
<td>4</td>
<td>73.5</td>
<td>Fair</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>Fair</td>
</tr>
<tr>
<td>7</td>
<td>76</td>
<td>Fair</td>
</tr>
<tr>
<td>8</td>
<td>76.5</td>
<td>Fair</td>
</tr>
<tr>
<td>9</td>
<td>77.5</td>
<td>Fair</td>
</tr>
<tr>
<td>10</td>
<td>77.5</td>
<td>Fair</td>
</tr>
<tr>
<td>15</td>
<td>78.5</td>
<td>Fair</td>
</tr>
<tr>
<td>30</td>
<td>79</td>
<td>Fair</td>
</tr>
<tr>
<td>50</td>
<td>79.5</td>
<td>Fair</td>
</tr>
<tr>
<td>100</td>
<td>79.5</td>
<td>Fair</td>
</tr>
</tbody>
</table>

Summarizing the meaning of these results:

Reliability increases dramatically from 2 to 3 scale points. It then increases upwards to seven categories, leveling off so that 100
categories is no more meaningfully reliable than a seven category scale. These results were replicated successfully in a later investigation by Preston & Colman (2000).

**Summary and Conclusions**

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). This scale is our “Gold Standard.”

   There is also an important caveat: Any Wine Scale (including The Three Stooges) 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.

   In summary, the purpose of this paper is to provide criteria for selecting wine scales that are not only useful to researchers, 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 exits 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 not overlapping. The scale also needs to have wide coverage. And finally, the Rater must be properly trained in order for the findings to be accurate. For a more comprehensive article on wine rating scales the interested wine researcher is referred to the work of Cicchetti & Cicchetti (2009):
References


