Summary 189

VITICULTURE & ENOLOGY
UNIVERSITY OF CALIFORNIA DAVIS

Evaluation of the consistency of wine quality assessments from expert wine tasters

By: R. Gawell and P. Godden

In: Australian Journal of Grape and Wine Research. 14(1): 1-8. 2008

- Food and beverage quality has been defined as "the ability of a set of inherent characteristics of a product to fulfill the requirements of customers". Evaluation of wine quality is normally undertaken by "wine experts" - tasters whose experience and training allows them to evaluate whether defects are present and whether the wine typifies the variety, region, or style intended. But, as the authors note, this is no guarantee that individual experts will weigh the different dimensions of a wine in a similar way.

- These authors believe that, for a quality score to start having any value, an expert taster should first be able to demonstrate her/his ability to reproduce a quality score when assessing the same wine several times. The goal of this study was to find out if expert tasters could indeed do that.

- The authors collected wine quality score data from 571 wine experts over a period of 15 years. (The experts were participants in an "Advanced Wine Assessment Course" conducted by the Australian Wine Research Institute). The wine expert demographics were as follows: 75% winemakers, 14% wine traders, 8% wine researchers, and 3% wine journalists. As for the wines, they represented a diverse range of varieties and styles familiar to the group, including Chardonnay, Sauvignon blanc, Riesling, and their blends, for whites; and Syrah, Cabernet Sauvignon, Grenache, Pinot noir, Merlot, and their blends, for reds.

- During the study, test wines and their duplicates were embedded within large flights of wines which represented particular varieties and styles each session. Thus, even though the expert tasters were aware that their judging performance was being evaluated, they were unaware which of the presented wines were being used to evaluate their consistency. The judges scored the wines for "overall quality" using a 20-point scale. In arriving at this score, judges were allowed to weigh the different aspects of wine quality in any way that they considered fit. Once the data was collected, the authors used a variety of sophisticated statistical analyses to evaluate: 1) each individual wine expert's consistency, 2) the individual wine expert's ability to discriminate wines, 3) intra-panel consistency of small groups of 3 experts, 4) the consistency between the assessment of red and white wines, and 5) the evolution of a wine expert consistency over time.

- The authors make a distinction that is worth mentioning. Even though "reliability", "consistency" and "category agreement" might all seem the same to you (they certainly do to me), the authors did use different tests to measure each. Let's see how they defined each.
  They would consider a judge to be "reliable" if, the second time around, his/her bottom wines still got low scores, and his/her top wines got proportionally high scores, even if the scores were all 2 points higher, or lower, the second time. But this judge would not be "consistent". To try to differentiate between these two, the authors measured 2 statistics ("regression correlation" to gauge reliability; and "absolute difference between scores", to measure consistency), but because even these two statistics were insufficient, they had to introduce a third statistic ("category agreement", or the percentage of times two scores for the same wine fell in the same quality category). In any event, let's see the results.

- **Results. 1) Consistency of individual assessors**.When the authors measured *reliability* - or the "regression coefficient" between both sets of scores –, they found that two thirds of judges showed significant reliability when judging red wines, and only half showed significant reliability with white wines. When the authors measured *consistency* – or the "absolute difference between scores"- ,they found that most of the judges were consistent. Finally, when the authors measured "category agreement", they found that it was moderate to very high, even though it was significantly higher for reds than for whites. In brief, **judges were better at reproducing a quality score for a red wine than for a white wine.** The authors attribute this difference to the possibility that, in red wines, the judges may have used visual color as a cue to quality. Color intensity has previously been shown to correlate well with flavor and other positive characteristics of wine.

- **2) Ability of individual assessors to discriminate wines**. By comparing the intra-wine and inter-wine score variability, the authors noticed that 64% of the **judges discriminated among red wines better than they did among white wines**. Overall, these results indicated that the majority of the judges achieved a consistent scoring pattern not by scoring all the wines in a narrow range, but rather by adequately discriminating among them and giving them widely different scores. Previous studies have suggested that the simultaneous presence of high reliability and high discrimination characterizes an experienced and confident judge.

- **3) Consistency of panels of assessors**. The authors observed that **red wine score consistency was improved when using the combined scores of 3 expert tasters**. Even if this trend was less clear for white wines, the authors believe these findings justify the current practice in Australian wine judgings of using a small panel of tasters.

- **4) Changes in assessor consistency over time**. Since this study spanned 15 years, it would be conceivable that the ability of the judges to give quality scores might have changed over time. When the authors tested this, they could see that no systematic change in performance had taken place. [*Does this mean that the average quality of Australian wine has been stagnant for the last 15 years?! Alternatively, an improvement in high-end categories may have been offset by the emergence of low-price-point brands. We don't have enough information*].

In conclusion, when judging "overall quality" expert tasters were better able to allocate red wines to the same quality category than white wines, and this ability improved by combining the scores of three wine experts. This indicates that the common practice of using a small panel of wine experts to judge wine "overall quality' is well-justified.

Author: Bibiana Guerra, Editor: Kay Bogart. This summary series funded by J. Lohr Vineyards & Wines.