# Quality Certification under Uncertainty: An Analysis of Wine Competition Ratings[*]

Gianni De Nicolò[†]
Johns Hopkins University
Carey Business School

Magalie Dubois[‡]
Burgundy School of Business

October 2025

**Abstract**

Wine competitions serve as certifiers of quality by awarding medals that influence consumer choice and producer strategies. Yet current rating protocols generate uncertainty from two sources: *rating risk*, due to variation in jury assignments and score aggregation within competitions, and *competition risk*, arising from inconsistent standards across events. We propose a standardized rating system that addresses rating risk by normalizing judges' scores and grouping wines into statistically significant quality classes. Using data from a wine competition, we show that this method reduces score variance and stabilizes award allocations. Extending the analysis, we use a model of intermediary certifiers to capture competition risk and show that heterogeneity in rating protocols diminishes the informational value of medals, raising consumer search costs and lowering producer profits. Standardization across competitions strengthens the credibility of ratings, improves welfare, and enables more efficient submission strategies for producers.

**JEL Classifications**: Q13, L15, C14.
**Keywords:** Wine ratings, Wine competition, ANOVA.

[*] This paper is an extensive revision of the paper previously titled "Which Wine Deserves a Medal? A Rating System for Wine Competitions" (AAWE Working Paper n.285). We thank the organizers of the Citadelles du Vin wine competition for having provided the 2022 competition data. We also thanks Robert Joseph, Jin Cao, Jean-Claude Ruf, and participants to the 3rd Annual Conference of the European Association of Wine Economists (EuAWE) and the 17th Annual Conference of the American Association of Wine Economists (AAWE) for their comments.

[†] Associate Professor of Practice, DipWSET, UC Davis Winemaking Certificate. Email: gdenico1@jhu.edu

[‡] Assistant Professor, Email: magalie.dubois@bsb-education.com

# 1 Introduction

Ratings provided by third-party organizations are ubiquitous in markets affected by adverse selection, as these entities serve as quality certifiers of goods and services. By reducing asymmetric information about product quality, ratings can improve welfare by lowering consumers' search costs and producers' costs of quality disclosure.

In wine markets, competitions act as certifiers by awarding medals that signal wine quality. These competitions are widely perceived as valuable marketing tools for wineries and represent profitable ventures for their organizers. Winning a medal is expected to enhance visibility and reputation among consumers and distributors, potentially leading to higher prices or increased sales. To improve their chances of having medals plastered on their bottles, wine producers often allocate a significant portion of their marketing budgets to submit wines to multiple competitions—partly explaining the recent proliferation of such events (Henrycs et al., 2016). While anecdotal evidence suggests that medals enhance marketing and revenue, systematic evidence on total revenue increases remains limited. A partial exception is Paroissien and Visser (2020), who found that medal-winning wines in French competitions command a 13% price premium. On the demand side, despite a growing literature on the economics of labels (Bonroy and Constantatos, 2014), there is limited research on the impact of medal labeling. Some evidence suggests that, despite consumer skepticism, medals influence wine purchase decisions (Neuringer et al., 2017).

Do the rating protocols used by wine competitions provide welfare-improving certification of wine quality? We address this question by examining current rating protocols and propose improvements aimed at enhancing consumer and producer welfare. Specifically, we design a rating system that reduces uncertainty in quality assessments and use a model of intermediary certifiers to evaluate the welfare benefits of standardizing rating protocols across competitions.

Current rating protocols typically follow three steps. First, each submitted wine is assigned to a jury of industry professionals who score their allocated wines on a 0–100 scale. Judges may report a single score or use a scorecard that breaks down quality factors, which are then summed to produce a total score. Second, a wine's final score is usually derived from a summary statistic of the judges' scores, such as the mean or median. Third, medals are awarded based on score intervals exceeding predefined thresholds, which may be constrained by a maximum fraction of wines eligible for medals. Medals are often categorized (e.g., Gold, Silver, Bronze).

The distribution of scores and medals varies significantly across competitions due to factors such as: 1) The number of submitted wines; 2) Jury allocation methods; 3) Number of judges per jury; 4) Use (or absence) of scorecards; 5) Criteria for medal eligibility. These variations introduce uncertainty, leading producers to solve a portfolio problem under uncertainty—balancing the expected payoff of winning a medal against the cost of unsuccessful submissions. Submitting to multiple competitions becomes a costly diversification strategy to mitigate the risk of not winning a medal.

Within a single competition, a wine's score may vary depending on the jury and the aggregation method used. The same wine may receive different scores depending on the jury to which it is assigned and the score aggregation method. Importantly, the score ranges established to grant a medal are not based on an assessment of statistically significant wine score differences. For example, a wine scored 89 may not receive a medal if the threshold is 90, despite the lack of a statistically significant difference between the two scores. We refer to this uncertainty as **rating risk**. Across competitions, differences in rating protocols introduce further uncertainty about expected scores. As shown by Hodgson (2009) examining the results of 13 US wine competitions, the probability of winning a medal at one competition is stochastically independent of the probability of winning a medal at another competition, suggesting that the award of a medal is primarily random. Much of the literature on wine evaluation has focused on judges' inconsistencies and lack of consensus (Bodington, 2020; Hodgson and Cao, 2014; Hodgson, 2008). However, even with perfect consensus and discriminatory ability, a flawed rating system design can undermine the informational value of ratings. We refer to this uncertainty as **competition risk**.

### Rating risk

Our proposed rating system builds on statistical methods used in food science (Lawless and Heymann, 2010), as applied to wine evaluations (Jackson, 2020; Lesschaeve and Noble, 2022), and originally developed by Amerine and Roessler (1983). To reduce rating risk, we introduce two key design features: (a) Standardization of judges' scores, and (b) Partitioning of scores into ranked, disjoint, quality-equivalent rating classes.

We apply a two-level standardization. First, we standardize each judge's scores using Z-scores to account for individual scaling differences while preserving rank order. This transforms each judge's score distribution to have zero mean and unit variance. Second, we re-scale these standardized scores relative to the overall distribution across juries, ensuring comparability across the full sample of wines.

Partitioning into quality-equivalent classes is based on Analysis of Variance (ANOVA), following De Nicolo (2024). We compute the Minimum Significant Difference (MSD) at a standard confidence level for each jury's rated wines. MSD units serve as a "numeraire" to convert standardized scores into Quality Values (QVs). Wines with the same QV belong to the same quality class. Since MSDs and QVs vary by jury, each jury effectively produces its own "price system." We determine a unified MSD—an "exchange rate"—that ensures the number of wines in each rating class does not exceed the competition's medal quota. This QV-targeting procedure assigns awards (Gold, Silver, etc.) based on statistically significant score differences.

Using data from an actual wine competition, we show that our system reduces rating risk by lowering score variance through standardization, and by reducing variability in rating classes via averaging the exchange rate used to convert scores into quality classes.

### Competition Risk

Competition risk stems from uncertainty in the quality of ratings issued under different protocols. Using the model by Hopenhaym and Saeedi (2022), we assess the welfare implications of a competitive equilibrium involving consumers, producers, and multiple wine competitions under the assumption that consumers and producers observe wine ratings but cannot distinguish between rating protocols used by different competitions.

Our model yields two key results. First, heterogeneity in rating protocols reduces the informational value of ratings, increasing consumers' search costs and lowering producers' expected profits. Second, standardization improves welfare by making ratings comparable across competitions. This enhances the reputation of competitions, helps consumers make informed choices, and enables producers to better plan submission strategies and production choices.

**Literature**

Our paper contributes to the literature on certification and quality disclosure. Early work reviewed by Dranove and Jin (2010) examined how competition in product and certification markets affects information disclosure. Empirical studies span sectors such as healthcare, education, finance, and consumer goods. Building on progress in information theory (Ganuza and Penalva, 2010; Kentszow and Kamenica, 2016), recent research explores how rating system design affects welfare depending on market conditions (e.g., Stahl and Strautz, 2017; Bizzotto and Harstad, 2023; Hopenhaym and Saeedi, 2024). Empirical studies have examined certification in eBay (Elfenbein et al., 2015; Saeedi, 2019, Hui et al., 2023), Medicare Advantage (Hopenhaym and Saeedi, 2022, Vatter, 2025), and online labor markets (Apostolos et al., 2019). Our examination of the welfare properties of the standardization of rating protocols' focuses on the quality of ratings provided by multiple wine competitions.

We also contribute to the literature on expert ratings in wine markets, as reviewed by Dubois et al. (2025), and specifically on rating aggregation. Balinski and Laraki (2010, 2017) analyzed aggregation rules and proposed using the median to reflect "majority judgment". They showed that the median is a statistic of an aggregation function of judges' scores consistent with a set of basic set of preference axioms, assuming experts share a common language—analogous to judges using a standardized scorecard. Modeling expert scores as capturing "true" quality and a bias component, Cao and Stokes (2010) and Cao (2014) modeled expert scores as a combination of true quality and bias, while Cicchetti (2009), Gergaud et al. (2021) and Carayol and Jackson (2024) proposed aggregation methods that account for bias variation. Our approach differs from these contributions since we introduce statistical significance of score differences to generate rating classes—a feature not addressed in prior literature. De Nicolo (2024) presents a model for generating ranked, disjoint quality classes based on significant score differences; we extend this to multiple juries in the context of wine competitions.

The remainder of the paper is organized as follows: Section 2 reviews rating protocols of major wine competitions. Section 3 details the architecture of our proposed rating system. Section 4 implements the system using data of a medium sized wine competition. Section 5 presents a model to assess the benefits of rating protocol standardization and offers best-practice recommendations. Section 6 concludes.

## 2 A brief overview of wine competitions

The *Global Wine Medal Rating* (2022) report (GWMR henceforth) summarizes information about the awarded wines taken from more than 250 of the largest wine competitions out of at least 600 held in the world every year. Figure 1 reports statistics of worldwide wine competitions in the GWMR database. Europe is the region of the world where the largest number of national and international competitions are held, followed by competitions in North America. In almost every country, several regional wine competitions take place.

Figure 1: **Wine Competitions Worldwide (GWRM 2020-2021)**

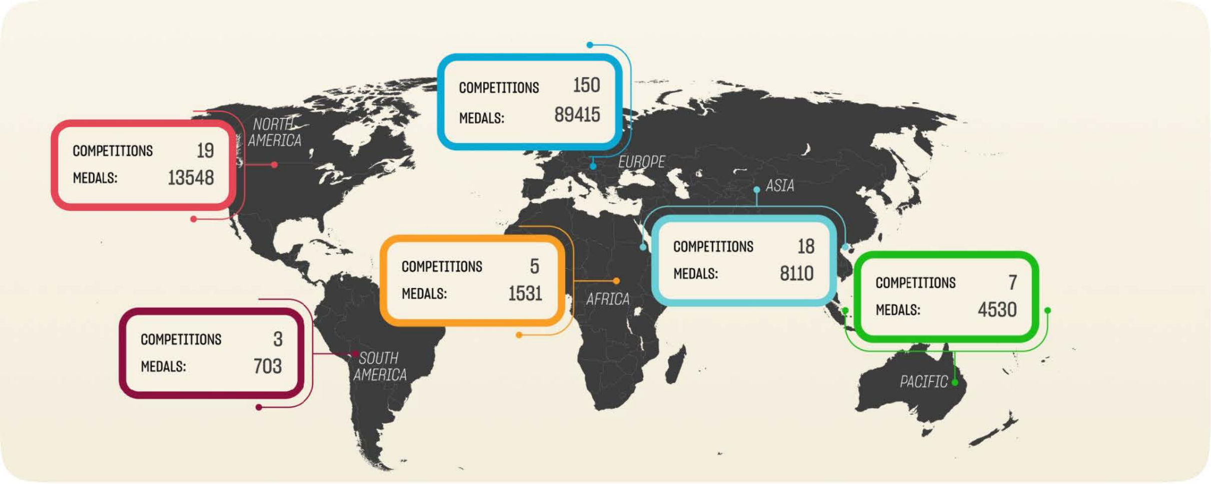


Table 1 reports data on entries, medal awards, medal categories, and score ranges for the top five international competitions identified by GWMR. The percentage of medals awarded varies considerably across competitions, with the highest rates observed in DWWA, ISWC, and IWC. However, medal rating categories are not directly comparable. For example, MV and CMB award a gold medal to wines scoring 85+ points, while DWWA assigns only a Bronze for that score. MV requires 90 points for a gold medal, whereas IWC classifies wines with that score as Silver. A Grand Gold award is given for wines scoring 95+ at MV, but only 92+ at CMB. As noted by *Meininger's International* (2023), these differences in rating classifications complicate producers' understanding of how their wines will be evaluated and make it difficult for consumers to interpret the quality value of wines across competitions.

Table 1: **Top five International Competitions: Medals and Score Ranges**

| **Competition** | **Avg. Entries** | **Medals % of entries** | **Medal Categories and Score Ranges** |
|---|---|---|---|
| DWWA | 18,500 | 82% | Bronze: 86–89, Silver: 90–94, Gold: 95–96, Platinum: 97–100 |
| MV | 12,000 | 30% | Silver: 85-89, Gold: 90-94, Grand Gold: 95-100 |
| IWC | 7,000 | 71% | Bronze: 85–89, Silver: 90–94, Gold: 95–100 |
| IWSC | 12,000 | 83% | Bronze: 75–79, Silver: 80–89, Gold: 90–100 |
| CMB | 10,000 | 30% | Silver: 85-87.9, Gold: 88–91.9, Gran Gold: 92-100 |

Source: Wine Competitions websites.
Decanter World Wine Awards (DWWA), UK; Mundus Vini (MV), Germany; International Wine Challenge (IWC), UK; International Wine & Spirit Competition (IWSC), UK; Concours Mondial de Bruxelles (CMB), Belgium. Average entries and % of medals data obtained from Internet search on Competitions and Wine Industry reports during the period 2021-2025.

Table 2 outlines the general protocols used by these competitions to select and assign wines to judges. In all cases, judges evaluate wines blindly but are informed of the wine category. DWWA additionally provides juries with the price range of the wines. Furthermore, DWWA, IWC, and IWSC allow for score adjustments through re-evaluation by jury co-chairs or panels, although the criteria for such re-evaluations are not formally defined. MV and CMB follow the rating protocol established by the *Organisation Internationale de la Vigne et du Vin* (OIV).

Table 2: **Top five International Competitions: Scoring Methodology**

| Competition | Scoring Methodology |
|---|---|
| DWWA | Wines are blind tasted in regional flights with contextual information (grape, vintage, price band). Scores are finalized through panel discussion. Silver and Gold medals are re-tasted by senior judges and Co-Chairs. |
| MV | Wines are blind tasted. The rating protocol follows OIV (2025). |
| IWC | Wines are blind tasted in flights grouped by style and region. Co-Chairs re-taste all medal candidates. |
| IWSC | Wines are blind tasted. Scores are discussed in panels. |
| CMB | Wines are blind tasted and Wines are grouped by type. The rating protocol follows OIV (2025). |

The OIV (2025) rating protocol includes the following rules:

- Scores are determined using a wine scorecard based on *International Union of Oenologists* standards, requiring individual ratings of ten quality factors. These are aggregated on a 0–100 scale.

- Each wine is rated based on the average of judges' scores, excluding scores that deviate by more than seven points from the average.

- The jury president may request a second evaluation by another jury, with the second jury's scores serving as the final rating.
- The total number of medals awarded must not exceed 30% of submissions. If this threshold is exceeded, samples with the lowest scores are eliminated.

The heterogeneity of rating protocols introduces substantial uncertainty regarding how wines are evaluated and complicates the interpretation of ratings across competitions. Introducing a rating system that mitigates this uncertainty and makes protocols—and thus medal values—broadly comparable would improve the informational quality of wine ratings.

# 3 A rating system for wine competitions

A wine competition is a set of $P$ juries ($p \in \{1, .., P\}$), each composed of $N_p$ different judges ($i \in \{1, .., N_p\}$), who evaluate $M_p$ wines ( $j \in \{1, .., M_p\}$) according to a set of quality "factors" indexed by $q \in \{1, .., Q\}$ on a numerical rating scale. The raw score of quality factor $q$ of wine $j$ of judge $i$ in jury $p$ is denoted by $X_{ij}^q(p)$. The raw total score of wine $j$ by judge $i$ in jury $p$ is the weighted sum of judge $i'$'s quality factor scores, given by $X_{ij}(p) = \sum_{q=1}^{Q} w_q X_{ij}^q(p)$, where $w_q \in (0, 1)$ is the weight of quality factor $q$, with $\sum_{q=1}^{Q} w_q = 1$. In this paper, the allocation of wines to juries is treated as given[1].

We aggregate scores by averages. If the distribution of judges' scores is approximately normal, then the mean and the median are approximately equal, implying that the use of average scores is consistent with Balinski and Laraki's "majority judgment" criterion. The aggregate raw score of wine $j$ in jury $p$ is the mean of judges' scores of wine $j$, given by $X_j(p) = N_p^{-1} \sum_{i=1}^{N^p} X_{ij}(p)$

## 3.1 The R distribution

The total number of wines in the competition is $N = pM_p$. The distribution of raw aggregate score is called the *R distribution*. Re-indexing wines by $k \in \{1, 2, ...., N\}$, the location $\mu_R$ (mean)

[1] The experimental design of the standard wine competition is an incomplete block design, where in each block (jury) different treatments (judges) are applied to different responses (wine scores). While it is not feasible to let all judges evaluate a large number of wines, other designs might be used to improve the determination of significant score differences. This is an important topic that is left to future research.

and scale $\sigma_R$ (standard deviation) of the R distribution are given by:

$$\mu_R = N^{-1} \sum_{k=1}^{N} X_k \equiv N^{-1} \sum_{k=1}^{N} \left( N_p^{-1} \sum_{i=1}^{N_p} \sum_{q=1}^{Q} w_q X_k^q \right) \tag{1}$$

$$\sigma_R = \sqrt{\frac{1}{N} \left( X_k - N^{-1} \sum_{k=1}^{N} \left( N_p^{-1} \sum_{i=1}^{N_p} \sum_{q=1}^{Q} w_q X_k^q \right) \right)^2} \equiv \frac{1}{\sqrt{N}} \sqrt{\sum_{k=1}^{N} \left( X_k - \mu_R \right)^2} \tag{2}$$

Equations (1) and (2) show that the location and scale of the R distribution depend on the number of wines in the competition, the allocation of wines to jury, the number of wines evaluated by each jury, the number of judges in each jury, the different ways judges use the wine scorecard, and the chosen set of weight $(w_q)_{q=1}^{Q}$ of a wine scorecard. Differences in these factors across wine competitions will determine an unavoidable uncertainty on the probability of winning a medal due to the differences arising from both the sample of wines submitted to each competition, as well as due to different protocols to score wines and assign medals of each wine competition.

## 3.2 The S distribution

The two-level standardization of raw aggregate wine scores we adopt delivers a distribution of standardized wine scores called the *S distribution.*

The level-one standardization addresses the issue of comparability of judges' scores of each jury. This comparability is established by a standardardization that makes the location and scale of each judge's score distribution the same. The aggregate raw score of each judge is standardized with respect to the location and scale of his/her score distribution using a Z-score. As mentioned in the introduction, this transformation takes into account the heterogeneity with which each judge may use the aggregate scales of the competition's wine scorecard due to tasting ability and idiosyncratic preferences, while preserving a judge's wine rankings. Formally, the level-one standardized score of wine $j$ evaluated by judge $i$ of jury $p$ is given by :

$$Z_{ij}(p) = \frac{X_{ij}(p) - \mu_i(p)}{\sigma_i(p)} \tag{3}$$

where $\mu_i(p) = M_p^{-1} \sum_j X_{ij}(p)$ is the mean and $\sigma_i(p) = \sqrt{M_p^{-1} \sum_j (X_{ij}(p) - \mu_i(p))^2}$ is the standard deviation of judge $i$'s aggregate raw scores. The Z-score of wine $j$ in jury $p$ is the average of the Z-scores of the judges in the jury, given by:

$$Z_j(p) = N_p^{-1} \sum_{i=1}^{N_p} Z_{ij}(p) \tag{4}$$

The level-two standardization addresses the issue of comparability of judges' scores across juries. Each jury evaluates different sets of wines which may differ by quality. Moreover, juries are composed of different judges who evaluate a different number of wines. The comparability of judges' evaluations across juries is accomplished by standardizing judges' Z-scores with respect to the location and scale of the entire distribution of raw aggregate scores. In taking location and scale of the entire distribution, we essentially smooth out differences of location and scale of each jury. Under the level-two standardization, the Z-score of wine $j$ in jury $p$ is standardized with respect to the location and scale of the R distrubution of wines' raw scores. The (doubly) standardized score of wine $j$ of jury $p$, denoted by $S_j(p)$, satisfies:

$$\frac{S_j(p) - \mu_R}{\sigma_R} = Z_j(p) \tag{5}$$

which yields:

$$S_j(p) = \mu_R + \sigma_R Z_j(p) \equiv \mu_R + \sigma_R N_p^{-1} \sum_{i=1}^{N_p} (\frac{X_{ij}(p) - \mu_i(p)}{\sigma_i(p)}) \tag{6}$$

The standardized score of wine $j$ of jury $p$ is a linear function with intercept $\mu_R$ and slope $\sigma_R$ multiplied by the average of standardized scores of the judges in jury $p$.

Re-indexing wines by $k \in \{1, 2, ...., N\}$, the location $\mu_S$ and scale $\sigma_S$ of the S distribution are given by:

$$\mu_S = N^{-1} \textstyle\sum_{k=1}^{N} S_k = \mu_R \tag{7}$$

$$\sigma_S = \sigma_R \tfrac{1}{\sqrt{N}} \sqrt{\textstyle\sum_{k=1}^{N} Z_k^2} \tag{8}$$

By construction, the location of the S distribution equals that of the R distribution. Comparing Equation (2) with Equation (8), the scale of the S distribution is smaller than that of the R distribution for a sufficiently large $N$. The lower scale of the S distribution relative to the R distribution can be viewed as reflecting the reduction of the risk that a wine is evaluated by a particular jury.

### 3.3 QV-based rating classes

The total number of medals and the rating classes establishing different quality medals based on the computation and selection of QVs is implemented in two steps.

The first step is to establish the maximum percentage $\tau$ of wines that will be awarded a medal. The choice of the percentage $\tau$ is of commercial relevance, as witnessed by the upper bound of 30% set by OIV rules, which limits the discretion of wine competitions to vary this percentage according to the number of submissions received. A reduction of uncertainty about the rating system adopted by a wine competition, and transparency considerations, would suggest that this parameter should be disclosed by a wine competition at the outset.

In the second step, the number of rating classes are established to allocate medals indexing different quality levels. A simple method is to assign different medals according the score percentiles of the S distribution of winners. Yet, the scores of the S distribution do not reflect scores' significant differences. We make the S distribution of winners conditional on significant score differences by computing wines' QVs, which deliver disjoint quality rating classes.

We compute the wines' QVs and derive rating classes by estimating the MSDs using the ANOVA results for each jury. We set the MSD equal to the *Fisher Least Significant Difference* (FLSD) at a 5% significance level. The FLSD is the smallest difference that exists between two significantly different score means. We treat the FLSD as a benchmark, since it is the most liberal test of pairwise comparisons, being based on a Type I error rate that assumes individual pairwise comparisons[2].

[2] Pairwise comparisons of means following statistically significant $F$-tests are used to detect which particular means in a group are significantly different. When multiple independent tests are conducted, each test has an inherent Type I error rate $\alpha$, but the overall *family-wise* Type I error rate accounting for all the $(n-1)/2$ comparisons is equal to $1-(1-\alpha)^n$, where $n$ is the number of comparisons. The FLSD is the most liberal, as it does not control for the family-wise error rate. In applications, the FLSD is considered "protected" from underestimation of the Type I error rate by an ANOVA $F$-test resulting in a very small p-value. For a review of multiple pairwise mean comparisons following ANOVA, see Sauders and De Mars (2019).

The FLSD of each jury $p$ is given by:

$$FLSD(p) = t_{(\alpha,df)}\sqrt{2v(p)/n(p)} \tag{9}$$

where $t_{(\alpha,df)}$ is the Student t-value at a significance level of $\alpha$, $df = M_p(n(p)-1)$ are the degrees of freedom, $M_p$ is the number of wines, $n(p)$ is the number of scores, and $v(p)$ is the error variance of the jury. The estimated FLSD(p) is treated as an *indivisible unit of account*, or "currency", of jury $p$. The QV of wine $j$ in jury $p$ is given by :

$$QV(S_j(p)) = INT\big(\frac{S_j(p)}{FLSD(p)}\big) \tag{10}$$

where the $INT$ operator transforms the score of a wine into an integer number. The QV automatically delivers *ranked quality rating classes* of wines evaluated by jury $p$ .

QVs differ across juries depending on the composition of the jury and the set of evaluated wines. As a result, differences of juries' currencies are reflected in different values of $FLSD(p)$. To compare QVs across juries, we need to use the same FLSD for the entire sample using an appropriate "exchange" rate. Let $FLSD(k)$ the currency of reference, where $k$ is a jury in the sample. The exchange rate of jury's currency $p$ relative to jury's currency $k$ is $e(p,k) = \frac{FLSD(p)}{FLSD(k)}$. Then:

$$QV^k(S(p)_j) = INT\big(\frac{S(p)_j}{FLSD(p)}e(p,k)\big) = INT\big(\frac{S(p)_j}{FLSD(k)}\big) \tag{11}$$

The currency conversion is immaterial with respect of wine rankings within each jury, i.e., any conversion in a particular currency is *rank-preserving.* What changes is the granularity of the rating classes: the larger the $FLSD(k)$, the less granular is the distribution of rating classes. Importantly, changes in granularity determine different allocations of wines to rating classes.

### 3.4 QV targeting

The FLSD can be chosen by two complementary *QV-targeting* methods. Method QV1 uses the mean FLSD across juries as a simple estimate of the location of the distribution of FLSD across juries. Wines are classified in medal categories, subject to the constraint that the number of winning wines does not exceed the maximum percent $\tau$ of wines deserving a medal. If such constraint is violated using the FLSD mean, then the FLSD is recomputed to derive the number of rating classes for which the total number of winners equal the maximum percentage $\tau$ of winners. Method QV2 uses this second step if the wine competition establishes that the number of winners is closest to the percentage $\tau$ of submissions.

Formally, QV-targeting is implemented as follows. Denote with with $N_W(\tau)$ the total number of winners, with $W$ the number of desired rating classes, with $FLSDm$ the average FLSD across juries, and with $N_w^{QV}(\tau)$ the number of wines in rating class $w \in W$. Using method QV1, if the sum of wines in the chosen rating classes is less than the maximum percentage of wines to deserve a medal, i.e. $\sum_{w \in W} N_w^{QV}(\tau) < \tau N$. then the procedure is complete.

If, instead, the constraint on the maximum number of wines deserving a medal is violated, i.e. $\sum_{w \in W} N_w^{QV}(\tau) > \tau N$, then the QV-targeting optimum value of FLSD, denoted by $FLSD^*$, is found by solving the following minimization problem:

$$\min_{FLSD \in [f_{min}, f_{max}]} \Big| \sum_{w \in W} N_w^{QV}(\tau) - N_W(\tau) \Big|. \qquad (12)$$

where $[f_{min}, f_{max}]$ is the range of estimated FLSDs across juries. The $FLSD^*$ is found using a simple search algorithm on a discretized set of FLSD values in the range $[f_{min}, f_{max}]$, as we show in our application.

The $FLSD^*$ minimizes the absolute distance between $N_W(\tau)$ and $\sum_{w \in W} N_w^{QV}(\tau)$ by identifying the partition in QV-based rating classes whose sum of wines is closest to the number of winners. If at the $FLSD^*$ optimum $\sum_{w \in W} N_w^{QV}(\tau) = N_W(\tau)$, then all winners are allocated to QV-based rating classes. If at the $FLSD^*$ optimum we instead obtain $\sum_{w \in W} N_r^{QV}(\tau) \neq N_W(\tau)$, then the number of winners is set equal to $\sum_{w \in W} N_w^{QV}(\tau)$ to avoid the arbitrary spit of winners and non-winners in wines classified in the same QV-based rating class. Method QV2 implements the optimization problem of Equation (14) at the outset.

### 3.5 Summary

Implementing our proposed rating system with the standard dataset of a wine competition is straightforward. First, all scores across juries are standardized, creating the S distribution. Second, the choice of the maximum percent $\tau$ of the winner is established. Third, the FLSDs of each jury is computed via ANOVA, and the distribution of FLSDs across juries is determined. Lastly, the value of the FLSD that delivers the QV-based rating classes is determined via the QV-targeting procedures, obtaining the QV-based distribution of quality medals among the medal winners.

## 4 Application

We detail the implementation of our rating system using data of the 2022 Citadelles du Vin wine competition (CdV henceforth). CdV is an International Vine and Wine Organization (OIV) sponsored wine competition created in 2000 in Bourg-sur-Gironde, France. It was the first French competition to become a member of the World Federation of Major International Wine and Spirits competitions (VINOFED) in 2014. Since its creation, the juries, mostly composed of international judges, have awarded 27% of the 23,074 tasted wines and spirits samples a grand gold, gold or silver medal. CdV follwows thre OIV (2025) rating protocol. We illustrate the properties of our rating system with no reference and/or comparison with the results obtained in the actual wine competition[3].

### 4.1 The CdV dataset

The CdV dataset employed in this study includes scores of 124 dry white wines and 340 dry red wines evaluated during the 2022 competition. The raw scores of each wine were obtained by aggregating sub-ratings of ten quality factors following the OIV scorecard. White wines were evaluated by nine juries, red wines were evaluated by 18 Juries. Each jury was composed of five judges.

[3] An earlier dataset of the same CdV wine competition was analyzed by Balinski and Laraki (2010), who distinguished between a ”judge-based” scoring system that relies on judges' aggregate scores of quality factors, compared to a ”criterion-based” system where the evaluation of each wine's quality factor is considered to be evaluated by different judges. In their terminology, our rating system is a ”judge-based” system.

Table 3 shows the CdV(OIV) scorecard, where the ten quality factors are classified in four categories with relevant weights: Visual, including V1 and V2 quality factors, has a 13% total weight; Nose, including N1-N3, has a 30% total weight; Taste, including T1-T4 quality factors, has a 44% weight total, and Overall Judgement has a 13% weight. Note that the scores of quality factors in each of the five "quality" buckets are scaled according to their weight in the total score.

Table 3: **CdV(OIV) Scorecard**

| **Category** | **Quality factors** | **Code** | **1** | **2** | **3** | **4** | **5** | **sum rows** | ***weights*** | **by category** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Visual** | **Clarity** | **V1** | 5 | 4 | 3 | 2 | 1 | **15** | *0.04* | **0.13** |
| | **Appearance** | **V2** | 10 | 8 | 6 | 4 | 2 | **30** | *0.08* | |
| **Nose** | **Cleaness** | **N1** | 6 | 5 | 4 | 3 | 2 | **20** | *0.06* | **0.30** |
| | **Intensity** | **N2** | 8 | 7 | 6 | 4 | 2 | **27** | *0.08* | |
| | **Quality** | **N3** | 16 | 14 | 12 | 10 | 8 | **60** | *0.17* | |
| **Taste** | **Cleaness** | **T1** | 6 | 5 | 4 | 3 | 2 | **20** | *0.06* | **0.44** |
| | **Intensity** | **T2** | 8 | 7 | 6 | 4 | 2 | **27** | *0.08* | |
| | **Persistence** | **T3** | 8 | 7 | 6 | 5 | 4 | **30** | *0.08* | |
| | **Quality** | **T4** | 22 | 19 | 16 | 13 | 10 | **80** | *0.23* | |
| **Overall** | **Overall judgment** | **OJ** | 11 | 10 | 9 | 8 | 7 | **45** | *0.13* | **0.13** |
| | **score (sum columns)** | | **100** | **86** | **72** | **56** | **40** | **354** | **1.00** | **1.00** |

## 4.2 Comparing the R and S distributions

The comparison of the raw scores and the standardized score distributions illustrates the implications of standardization for the ranking of wines and the identification of "medal winners".

Table 4 reports descriptive statistics of raw wine scores, and Figure 2 depicts histograms of the raw score distribution compared to the normal distribution. Average and dispersion of scores of the two samples are similar, with the median and mean being very close. As shown in Figure 2, however, both distributions exhibit "fat" left tails, due to a set of wines ranked significantly lower than the median. As a result, both distributions do not match normality for score percentiles lower than the median.

Table 4: **R distribution: descriptive statistics**

| | **Mean** | **Median** | **Std.dev** | **Min** | **Max** | **no. of wines** |
|---|---|---|---|---|---|---|
| **Whites** | 87.25 | 87.60 | 3.24 | 71.80 | 94.20 | 124 |
| **Reds** | 87.10 | 87.40 | 3.60 | 74.80 | 95.00 | 340 |

Figure 2: **R Distributions**

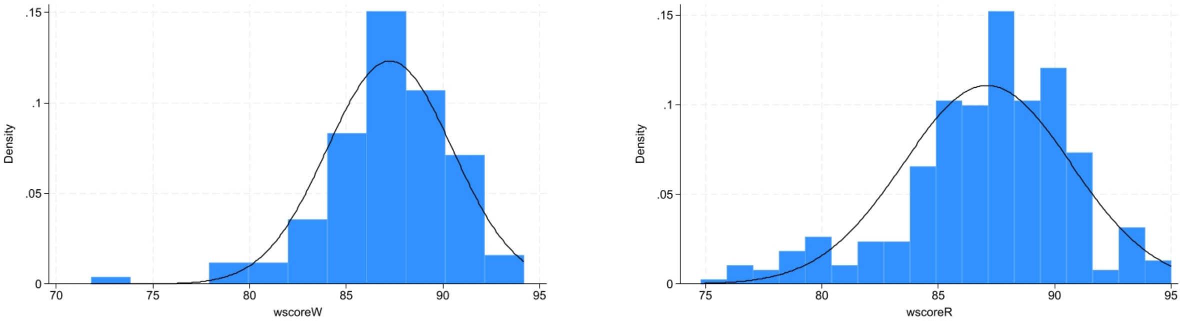


Table 5 reports descriptive statistics of the mean raw scores of juries and the range of the number of wines evaluated by each jury. Location and scale vary notably by jury due to the random assignment of wines and the different number of wines evaluated by each jury.

Table 5: **Descriptive statistics of mean raw scores across juries**

| **# of juries** | **Mean** | **std.dev** | **Min** | **Max** | **min # of wines** | **max # of wines** |
|---|---|---|---|---|---|---|
| 9 | 87.33 | 1.52 | 83.92 | 88.60 | 12 | 15 |
| 18 | 87.06 | 2.56 | 81.42 | 90.98 | 12 | 39 |

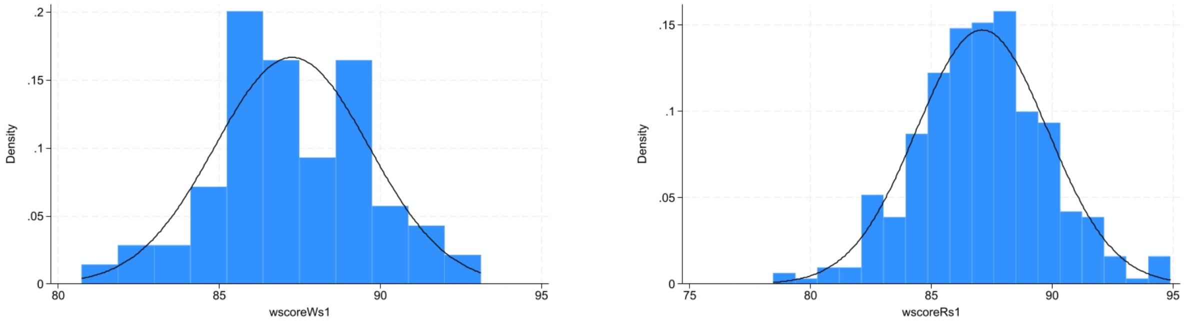
Turning to the S distribution, Table 6 reports the relevant descriptive statistics. Similarly to the raw score distribution, average and dispersion of standardized scores of the two samples are similar, with the median and mean being very close. However, the dispersion of the S distribution is remarkably lower than that of the R distribution, as witnessed by the lower standard deviation for both wine samples. Notably, the Spearman rank correlation coefficient between the R and S distributions is 0.88 for white wines and 0.70 for red wines: this means that the teo-level standardization of wine scores delivers important changes in the rankings of wines.

As shown in Figure 3, the S distribution does not exhibit ”fat” left tails, and appears for both white and red wines approximately normal.

Table 6: **S distribution: descriptive statistics**

| | **Mean** | **Median** | **Std.dev** | **Min** | **Max** | **Spearman**<br>**R-S distr.** | **no. of wines** |
|---|---|---|---|---|---|---|---|
| **Whites** | 87.25 | 86.99 | 2.39 | 80.73 | 93.11 | **0.88** | 124 |
| **Reds** | 87.10 | 87.06 | 2.71 | 78.45 | 94.89 | **0.70** | 340 |

Figure 3: **S distributions**



Would a wine producer prefer to have his/her submitted sample evaluated according to raw scores or standardized scores ? The estimate of the expected score per unit of score risk, given by the ratio of average score to the score standard deviation, is significantly larger for the S distribution, since averages are the same but the standard deviation of the S distribution is about 30 percent smaller than that of the R distribution. We can view the standard deviation of a distribution as reflecting in part the risk of a wine being assigned to a jury which evaluates a set of wines remarkably differently from the overall set of wines. A wine producer submitting a sample would be ex-ante better off if his/her sample is evaluated under the S distribution due to the reduction of this risk.

### 4.3 Medal rankings under the R and S score distributions

We compare the medal rankings of the R and S distributions unconditional on significant score differences. Without loss of generality, we assume that medals are granted to no more than the best 25% of wines, and three ”medals” are granted: Gold, for wines whose score is greater than the 95% percentile of the score distribution; Silver, for wines whose score is less than or equal than the 95% percentile, and greater than the 90% percentile of the score distribution; and Bronze, for wine whose scores are less than or equal to the 90% percentile of the score distribution and greater than the 75% percentile of the score distribution.

As shown in Table 7, the total number of winners under each score distribution is the same by construction, but the identity of the winners is different both within the set of winners and within ranking classes. Considering white wines, 8 wines that are R-distribution winners are not S-distribution winners, and 8 wines that are S-distribution winners are not R-distribution winners. In other words, moving from the R to the S distribution, there are 8 winners’ ”exits” and 8 winner ”entrants”. Considering red wines, the differences in medal rankings are even starker: 28 wines that are R-distribution winners are not S-distribution winners, and 25 wines that are S-distribution winners are not R-distribution winners, resulting in 25 winners’ ”exits” and 28 winner ”entrants”. These differences in rankings are undoubtedly of commercial significance.

Table 7: **Medal winners under the R and S distributions**

| | | | **Medals** | | | **medal winners** | **medal winners** |
|---|---|---|---|---|---|---|---|
| | **distribution** | **total medals** | **Gold** | **Silver** | **Bronze** | **in R but not in S** | **in S but not in R** |
| **WHITES** | **R** | 31 | 6 | 3 | 22 | 8 | |
| | **S** | 31 | 6 | 6 | 19 | | 8 |
| **REDS** | **R** | 77 | 17 | 11 | 49 | 28 | |
| | **S** | 85 | 17 | 17 | 51 | | 25 |

Furthermore, the allocation of medals among winners is different. For white wines, the R-distribution gold, silver and bronze medals are 6, 3, and 22 respectively, while the S-distribution gold, silver, and bronze medals are 6, 6, and 19 respectively. For red wines, the R-distribution gold, silver and bronze medals are 17, 11, and 49 respectively, while the S-distribution gold, silver, and bronze medals are 17, 17, and 51 respectively. Due to the difference in the identity of total winners of the two distributions, the identity of winners of each medal group is also different and likely commercially significant.

## 4.4 QV-based rating classes

Using the S distribution, we compute the FLSD of each jury at a 5% significance level according to Equation (9). Note that the reliability of tests of mean differences using $F$-tests based on ANOVA rests on the assumptions of equality of variances across judges’ scores and approximate normality of the distribution of the relevant regression errors. Equality of variance across judges’ scores is guaranteed by the Z-score standardization of judges’ evaluations. We test the normality of the residuals associated with the ANOVA regression using Bera et al. (2016) test and the relevant computation of Quantile-to-Quantile plots (QQ plots) with 95% confidence intervals. We find that for each jury, the score distribution is inside the 95% confidence intervals of the

QQ plots, indicating that approximate normality is not rejected[4].

Table 8 reports the FLSD estimates. We highlight three results. First, there is a notable variation of FLSDs across juries, reflecting the random assignments of wines to juries, the difference of experts' evaluations in each jury, and the differences in the number of wines evaluated by each jury. Second, the FLSD estimates of each jury exceed one, which implies that a simple raw rating scale that identifies a wine better than another by a one-unit score difference is inconsistent with a statistically significant difference. Third, the mean FLSD is about 2 for both samples. This implies that, for example, a 91-point score is not significantly better than an 89.5-point score, since the difference of their score is not statistically significant.

Table 8: **FLSDs by Jury**

| **WHITES** | | **REDS** | | | |
|---|---|---|---|---|---|
| **jury** | | **jury** | | **jury** | |
| **W1** | 2.08 | **R1** | 2.08 | **R10** | 2.05 |
| **W2** | 1.72 | **R2** | 2.43 | **R11** | 1.87 |
| **W3** | 1.39 | **R3** | 1.65 | **R12** | 1.68 |
| **W4** | 1.99 | **R4** | 2.72 | **R13** | 1.66 |
| **W5** | 2.11 | **R5** | 1.73 | **R14** | 1.66 |
| **W6** | 1.71 | **R6** | 2.55 | **R15** | 2.56 |
| **W7** | 2.39 | **R7** | 2.51 | **R16** | 2.16 |
| **W8** | 2.01 | **R8** | 2.04 | **R17** | 2.12 |
| **W9** | 2.05 | **R9** | 1.77 | **R18** | 1.85 |
| **Mean** | **1.94** | **Mean** | **2.06** | | |
| **Std. dev.** | **0.29** | **Std. dev.** | **0.36** | | |
| **Min** | **1.39** | **Min** | **1.65** | | |
| **Max** | **2.39** | **Max** | **2.72** | | |

As mentioned in the previous section, the choice of using the same FLSD for all wines, as required by expressing the quality value of a wine in the same "currency", determines the granularity of the QV-based rating classes. The implications of chosing different values of the FLSD are illustrated in Table 9, which shows the QVs and the distribution of rating classes in ascending order with the number of wines in each class, for the minimum, the mean, and the maximum of the FLSD across juries. As the value of the FLSD decreases, the number of rating classes increases, leading to a more granular partition of wines within and across ratings. When we look at the first three highest QV-based rating classes (indicated in boldface), we see that the

[4] As shown in De Nicolo' (2024), if the Bera et al. (2016) test rejects the null of approximate normality, one can use a *robust* ANOVA, implemented by computing a modified $F$-test using trimmed means and winsorized variances. In this case, the assessment of significant mean differences is based on the Yean statistics, as detailed in Wilcox (2022).

number of medal winners declines with the use of lower FLSD values. The quality classification of wines by medal is also changing increasing the tighteness of the medal assignment criterion: for example, a wine that wins a medal under both rating classes obtained with values of $FLSD_1$ and $FLSD_2$, where $FLSD_1 > FLSD_2$, might win a silver medal under $FLSD_1$, or a bronze medal under $FLSD_2$.

Table 9: **QV and rating classes**

| | **FLSD max** | | **FLSD mean** | | **FLSD min** | |
|---|---|---|---|---|---|---|
| | **QV** | **wines** | **QV** | **wines** | **QV** | **wines** |
| **WHITES** | 27 | 17 | 41 | 2 | 58 | 2 |
| | 26 | 44 | 42 | 5 | 59 | 5 |
| | **25** | **48** | 43 | 16 | 60 | 9 |
| | **24** | **12** | 44 | 42 | 61 | 28 |
| | **23** | **2** | **45** | **32** | 62 | 30 |
| | | | **46** | **21** | 63 | 15 |
| | | | **47** | **6** | **64** | **24** |
| | | | | | **65** | **6** |
| | | | | | **66** | **5** |
| **no. of classes** | **5** | | **7** | | **9** | |
| **REDS** | 28 | 2 | 38 | 3 | 47 | 2 |
| | 29 | 5 | 39 | 11 | 48 | 3 |
| | 30 | 41 | 40 | 36 | 49 | 10 |
| | 31 | 122 | 41 | 89 | 50 | 29 |
| | **32** | **123** | 42 | 108 | 51 | 58 |
| | **33** | **37** | 43 | 62 | 52 | 90 |
| | **34** | **10** | **44** | **22** | 53 | 71 |
| | | | **45** | **7** | 54 | 48 |
| | | | **46** | **2** | **55** | **19** |
| | | | | | **56** | **6** |
| | | | | | **57** | **4** |
| **no. of classes** | **7** | | **9** | | **11** | |

We complete the application of our rating system by generating the final QV-based allocation of medals using QV-targeting. The results of the QV1 method are already depicted in Table **??**. Consider the third column corresponding to the FLSD mean. For white wines, the first three rating classes include 6, 21, and 32 wines in descending order. The total of medals would be 59, which is well above the maximum medals of 25% of the sample, equal to 31 wines. In this case, the second part of the QV1 method would be applied, as shown below. For red wines, the first three rating classes include 2, 7, and 22 wines in descending order. The total of medals is 31, which is well below the maximum medals of 25% of the sample, equal to 85 wines.

To illustrate the QV2 method, we apply QV-targeting to the number of medal winners identified by the wines with standardized scores greater than the 75% percentile of the S distribution. Table 10 report the results of the optimization problem defined in Equation (14), comparing the

QV-based rating classes to the rating classes based on the percentiles of the S score distribution. The optimal $FLSD^*$ attains a total number of QV ranked wines exactly equal to the number of wines with scores greater or equal than the 75th percentile of the S distribution.

Compared to the S distribution allocation of medals, the QV-based allocation assigns highest quality medals more sparingly than the allocation based on the percentiles of the S distribution. Consider white wines: only one gold medal is granted under QV-based rating, compared with six under the rating based on the percentiles of the S distribution, implying that five wines of that distribution are moved down one notch to silver. Likewise, a number of wines which are granted a silver medal under the percentiles of the S distribution are downgraded to bronze. The results for the red wines sample are similar: only six gold medals are granted under QV-based rating, compared with 17 under the rating based on the percentiles of the S distribution, implying that 11 wines of that distribution are moved down one notch to silver, and 11 wines that were silver medal winners are now only bronze metal winners..

Table 10: **Medal winners under the S and the targeted QV distributions**

| | | | | **Medals** | | |
|---|---|---|---|---|---|---|
| | $FLSD^*$ | **distribution** | **total medals** | **Gold** | **Silver** | **Bronze** |
| **WHITES** | | **S** | 31 | 6 | 6 | 19 |
| | **1.65** | **QV** | 31 | 1 | 8 | 22 |
| **REDS** | | **S** | 85 | 17 | 17 | 51 |
| | **2.40** | **QV** | 85 | 6 | 17 | 62 |

Summing up, the allocation of medal winners to quality ranked medals is more exacting under the QV-based rating due to the incorporation of significant score differences in the assignment of medals. A more stringent criterion of wine quality assessment in a wine competition might be preferred by a wine producer who invests significant effort and resources to make his/her wine objectively distinguishable from the ”crowd”.

## 5 Ratings as information structures

Wine ratings can be viewed as signals that provide decision-makers with information about the hidden quality of wines. In this section, we use the industry model developed by Hopenhayn and Saeedi (2023) to illustrate how the quality of information conveyed by ratings from multiple wine competitions affects welfare, measured by total surplus—the sum of consumer and producer surplus. The set-up of the model is as follows.

**Wine producers**. There is a unit mass of wine producers indexed by wine quality $z$ distributed according to an unobserved continuous cumulative distribution function (cdf) $F(z)$ on the score range $[0, 100]$. The cost of producing wine quantity $q$ of any quality is the same for all producers, given by a cost function $c(q)$, with $c'(q) > 0$ and $c''(q) > 0$. Given the price of wine $p$, the supply function $S(q)$ satisfies $S'(q) > 0$.

**Wine consumers**. On the demand side, there is a mass $M$ of consumers who decide whether to purchase a wine based on preferences $U = z + \theta - p$, where $z$ is the wine's quality, and $\theta \geq 0$ is a taste parameter representing the consumer's preference for wine relative to an outside good. The parameter $\theta$ is distributed according to a continuous and strictly increasing cumulative distribution function $\Psi(\theta)$. The utility of not purchasing wine is normalized to zero. Wines are differentiated solely by their expected quality level $z$.

**Wine competitions**. Information about wine quality is provided by $n$ wine competitions, whose wine scores are signals with discrete distribution $G_i(z), i = 1, ., n$ over the fraction of wines $w_i$ rated by wine competition $i$, where $\sum_{i=1}^{n} w_i = 1$. We assume that all market participants have the same posterior information about the expected wine quality after receiving the set of signals from the wine competitions, represented by the *mixture* of distributions $G(z) = \sum_{i=1}^{n} w_i G_i(z)$.

Information about wine quality is provided by $n$ wine competitions, each issuing scores that serve as signals. These scores follow a discrete distribution $G_i(z)$, where $i = 1, \ldots, n$, over the fraction of wines $w_i$ rated by competition $i$, with $\sum_{i=1}^{n} w_i = 1$. We assume that all market participants share the same posterior beliefs about expected wine quality after observing the set of signals from the competitions. These beliefs are represented by the mixture distribution:

$$G(z) = \sum_{i=1}^{n} w_i G_i(z) \tag{13}$$

**Equilibrium and welfare**. Hopenhayn and Saeedi (2022) characterize the competitive equilibrium of this model and define total surplus as a value function $TS(G(z))$, which depends on the mixture of ratings $G(z)$.

**Information structures**. The distribution of ratings $G(z)$ is an *information structure* describing how ratings relate to the underlying latent wine quality. An information structure improves the payoffs of decision makers by providing more information. Formally, given a prior distribution over qualities, an information structure specifies a conditional distribution of ratings given the qualities. The "quality" of information is captured by the notion of *garbling.* A garbled signal is one that has been probabilistically transformed from a more informative one. Therefore, a less garbled distribution is considered to contain higher-quality information.[5]

Following Ganuza and Penalva (2010) and Gentzkow and Kamenica (2016), given a common prior $\hat{F}(z)$ over wine qualities, the rating structure $G(z)$ is the distribution of the expected posterior of wine quality. Any information structure can thus be represented as a garbling of the unobserved $F(z)$. Consider a decision maker payoff $P(G_1)$ under signal $G_1(z)$, and $P(G_2)$ under signal $G_2(z)$. According to Blackwell's (1953) ordering, for any decision maker's payoff, $P(G_1) \geq P(G_2)$ if $G_2$ is less informative than $G_1$, that is, $G_2$ is a garbling of $G_1$.

In the context of wine competitions, each competition provides a noisy and potentially biased signal of latent wine quality. Differences in rating protocols mean the same wine can receive different ratings across competitions, different medal categories, or no medal at all. If market participants are unable to distinguish the quality of ratings issued by different wine competitions, then they observe only the rating mixture $G(z)$. But mixtures of ratings are less informative than the best individual rating system due to the convexity of the value function with respect to the probability distribution over wine qualities. Hence:

$$TS(G(z)) \leq \max\{TS(G_1(z)), TS(G_2(z)), ..., TS(G_n(z))\} \tag{14}$$

*If rating quality is not distinguishable by consumers and producers, then total surplus is reduced relative to the most informative rating system.*

In other words, multiple competitions with rating protocols that are not comparable increase

[5] A simple example of information structure is as follows. Let $X \in \{0, 1\}$ be the true wine quality (wine score <90 =0, wine score $\geq$ 90=1) with prior $P(X = 1) = 0.5$. If $X = 1$, the signal $S$ equals 1 with probability 0.9 and 0 with probability 0.1; if $X = 0$, then $S = 0$ with probability 0.9 and 1 with probability 0.1. This conditional distribution $P(S|X)$, combined with the prior, defines the information structure.

garbling. In practice, the inability of consumers to distinguish between a medal granted by competition A vis a vis that granted by competition B reduces welfare. Likewise, producers seek to pursue any medal granted by any wine competition, increasing their marketing costs.

Now consider the opposite scenario. Suppose consumers and producers are able to perfectly discriminate the wine ratings issued by different wine competitions. This ability would stem from perfect knowledge of the rating protocols used by different wine competitions, which would allow quality comparisons of the ratings. Due to the convexity of the value function, the previous inequality is reversed:

$$TS(G(z)) \geq \max\{TS(G_1(z)), TS(G_2(z)), ..., TS(G_n(z))\} \tag{15}$$

*If rating quality is distinguishable by consumers and producers, then total surplus is increased.*

In other words, more precise information (i.e., less uncertainty) allows consumers and producers to make better choices based on more reliable, transparent, and relevant information.

In practice, perfect discrimination and quality assessment of rating protocols may be unattainable due to differences in submitted samples, jury composition, and judge identities. However, the quality of information provided by wine competition ratings could be improved through the standardization of rating protocols, enabling consumers and producers to better evaluate and compare ratings across competitions. In the language of information theory, standardization reduces garbling.

The OIV (2025) rating protocol for sponsored wine competitions exemplifies welfare-enhancing standardization. Applying our rating system to data from the CdV wine competition suggests that the following standardized rating protocol rules—presented as recommendations—could improve rating quality and facilitate comparisons across wine competitions:

- *Disclosure of the percentage of wines eligible for medals.*

  Revealing the actual proportion of medal-winning wines relative to total submissions could enhance the evaluation of ratings across competitions. As shown by Hui et al. (2023) using eBay data, "raising the bar" for awarding medals may lead to self-selection among applicants, extending the right tail of the score distribution and encouraging higher-quality wines to enter. Similar to observations by Apostolos et al. (2019) in the context of online labor markets, the absence of an ex-ante criterion for award assignment may result in rating inflation. This inflation can stem from competitions aiming to maximize submissions and award a large number of medals, thereby diluting the reputational value of the ratings.

- *Implementation and disclosure of a standardized method for judges' scores.*

  Our analysis shows that relying on raw score distributions risks under- or overrating wines due to the random assignment of wines to juries. While various standardization methods could be adopted, our proposed two-level standardization is both computationally simple and effective in mitigating risks associated with the randomness of jury assignment.

- *Adoption and disclosure of a method for defining rating classes based on statistically significant score differences.*

  We demonstrate that defining rating classes based on statistically significant score differences can alter medal allocations, even when standardized scores are used. Although multiple methods could be developed for this purpose, our QV-based approach relies on standard ANOVA techniques and is straightforward to implement.

## 6 Conclusion

Wine competitions play a central role in certifying wine quality, yet the lack of standardized rating protocols introduces significant uncertainty for both producers and consumers. This paper has shown that current rating systems generate two distinct types of risk: rating risk, arising from variability within a single competition, and competition risk, stemming from differences across competitions. These risks reduce the informational value of ratings, distort market signals, and hinder welfare-enhancing outcomes.

To address these issues, we proposed a statistically grounded rating system that standardizes judges' scores and partitions wines into quality-equivalent classes based on statistically significant score differences. This system reduces rating risk by improving score comparability and lowers competition risk by enabling consistent interpretation of ratings across competitions.

Using a modified model of intermediary certifiers, we demonstrated that standardization enhances both consumer and producer welfare. Consumers benefit from reduced search costs and clearer quality signals, while producers gain from more predictable submission outcomes and improved strategic planning. Moreover, standardization fosters reputation effects among competitions, encouraging better practices and more reliable certification.

Our findings suggest that wine competitions—and potentially other markets relying on expert ratings—can significantly improve their role as quality certifiers by adopting standardized rating protocols. Future research may explore the broader applicability of our framework to other domains where expert evaluations are central to market functioning.

## References

[1] Amerine, Maynard A., and Edward B. Roessler, 1983, *Wines: Their Sensory Evaluation*, W.H Freeman and Company, New York and San Francisco.

[2] Apostolos, Filippas, John J. Horton, and Joseph M. Golden, 2019, Reputation Inflation, NBER Working Paper 25857, May.

[3] Balinski, M., and Laraki, R., 2007, A theory of measuring, electing, and ranking. Proceedings of the National Academy of Sciences of the United States of America, 104(21), 8720–8725. https://doi.org/10.1073/pnas.0702634104

[4] Balinski, Michel, and Rida Laraki, 2010, Majority Judgment, The MIT Press, Cambridge, Massachusetts.

[5] Bera, Anil K., Antonio F. Galvao, Liang Wang, and Zhijie Xiao, 2016, A new characterization of the normal distribution and test of normality, *Econometric Theory*, Vol. 332, no.5: 1216-1252.

[6] Blackwell, David, 1953, Equivalent Comparisons of Experiments, *The Annals of Mathematical Statistics*, Vol. 24, No. 2: 265-272.

[7] Bitter, Christopher, 2017. Wine Competitions: Reevaluating the Gold Standard, *Journal of Wine Economics*, 124: 395–404.

[8] Bizzotto, Jacopo and Bard Harstad, 2023, The Certifier for the Long Run, *International Journal of Industrial Organization*, 87: 1-19.

[9] Bonroy, Olivier, and Christos Constantatos, 2014, On the Economics of Labels: HOw Their Introduction affects the Functioning of Markets and Welfare of all Participants, *American Journal of Agricultural Economics*, 97(1): 239–259.

[10] Carayol, Nicolas, and Matthew O. Jackson, 2024, Finding the wise and the wisdom in a crowd: estimating underlying quality of reviewers, *The Economic Journal*, 134: 2712-2745.

[11] Cicchetti, Domenic V., 2009, A Proposed System for Awarding Medals at a Major U.S. Wine Competition, *Journal of Wine Economics*, Volume 4, Issue 2, Winter: 242-247.

[12] Bodington, J. 2020. Rate the Raters: A Note on Wine Judge Consistency. Journal of Wine Economics, 15(4), 363–369. https://doi.org/10.1017/jwe.2020.30

[13] Cao, Jing, 2014, Quantifying Randomness Versus Consensus in Wine Quality Ratings, *Journal of Wine Economics*, Vol. 9, n.2: 202-213.

[14] De Nicolò, Gianni, 2024, Wine Ratings and Commercial Reality, *Journal of Wine Economics*, vol. 20, 1: 1-25.

[15] Dubois M., Georgantzis N., Cardebat J.M., 2025, External Evaluations under Quality Uncertainty: the Market for Wine Ratings, *Wine Economics and Policy*, Vol. 14(1): 131-162.

[16] Dranove, David, and Ginger Zhe Jin, 2010, Quality Disclosure and Certification: Theory and Practice, *Journal of Economic Literature*,48:4: 935–963

[17] Elfenbein, Daniel, Raymond Fisman, and Brian McManus, 2015, Market Structure, Reputation, and the Value of Quality Certification, *American Economic Journal: Microeconomics*, 7(4): 83–108.

[18] Ganuza, J.-J. and J. S. Penalva, 2010, Signal Orderings Based on Dispersion and the Supply of Private Information in Auctions, *Econometrica*, 78: 1007–1030.

[19] Gentzkow, M. and E. Kamenica, 2016, A Rothschild-Stiglitz Approach to Bayesian Persuasion, *American Economic Review*, 106(5): 597–601.

[20] Gergaud , O. , Ginsburgh, V. and Moreno-Ternero, J.D., 2021. Wine ratings: Seeking a consensus among tasters via normalization, approval, and aggregation, *Journal of Wine Economics*, vol. 16(3): 321–42.

[21] Global Wine Medal Ratings, 2022, Global Wine Medals and Competitions Report - 2020-2021, https://gwmr.gustos.life/reports

[22] Henryks, J., Ecker, S., Turner, B., Denness, B., and Zobel-Zubrzycka, H., 2016, Agricultural Show Awards: A Brief Exploration of Their Role Marketing Food Products, *Journal of International Food and Agribusiness Marketing*, 28(4), 315–329.

[23] Hodgson, R. T. 2008. An Examination of Judge Reliability at a major U.S. Wine Competition. *Journal of Wine Economics*, 3(2), 105–113.

[24] Hodgson, R. T. 2009. An analysis of the concordance at 13 US wine competitions. *Journal of Wine Economics*, Vol.4, n.1: 1-9.

[25] Hopenhayn, Hugo, and Maryam Saeedi, 2022, Optimal simple ratings, mimeo

[26] Hopenhayn, Hugo, and Maryam Saeedi, 2023, Optimal information disclosure and market outcomes, *Theoretical Economics*, Vol. 18: 1317-1344

[27] Hui Xiang, Maryam Saeedi, Giancarlo Spagnolo, and Steven Tadelis, 2023, Raising the Bar: Certification Thresholds and Market Outcomes, *American Economic Journal: Microeconomics*, 15(2): 599–626.

[28] Jackson, Ronald S., 2020, *Wine Science: Principles and Applications*, 5th Edition, Academic Press, Elsevier Ltd, Amsterdam.

[29] Lesschaeve, Isabel, and Ann C. Noble, 2022, Sensory Analysis of Wine, Chapter 7 in Andrew G. Reynolds Editor, *Managing Wine Quality. Volume I: Viticulture and Wine Quality*, 2nd Edition, Woodhead Publishing, Elsevier, North Holland.

[30] Lawless, Harry T., and Hildegarde Heymann, 2010, *Sensory Evaluation of Food: Principles and Practice*, 2nd Edition, Springer, New York.

[31] Meninger's International, 2023, The Confusing World of Wine Competitions, October 8, https://www.meiningers-international.com/wine/insights-wine/confusing-world-wine-competitions

[32] Neuninger, R., Mather, D., and Duncan, T., 2017. Consumer's skepticism of wine awards: A study of consumers' use of wine awards. *Journal of Retailing and Consumer Services*, 35, 98–105. https://doi.org/10.1016/j.jretconser.2016.12.003

[33] Paroissien, E., and Visser, M., 2020, The Causal Impact of Medals on Wine Producers' Prices and the Gains from Participating in Contests. *American Journal of Agricultural Economics*, 102(4), 1135–1153. https://doi.org/10.1002/ajae.12037

[34] Quandt, Richard E. 2006, Measurement and Inference in Wine Tasting, *Journal of Wine Economics*, 1(1): 7-30.

[35] Saeedi, Maryam, 2019, Reputation and adverse selection: theory and evidence from eBay, *RAND Journal of Economics*, Vol. 50, No. 4: 822–853.

[36] Sauder, Derek C, and Christine E. DeMars, 2019, An Updated Recommendation for Multiple Comparisons, *Advances in Methods and Practices in Psychological Science*, Vol. 2(1) 26–44.

[37] Stahl, Konrad and Roland Strausz, 2017, Certification and Market Transparency, *The Review of Economic Studies*, Vol. 84, No. 4: 1842-1868.

[38] Vatter, Benjamin, 2025, Quality Disclosure and Regulation: Scoring Design in Medicare Advantage, *Econometrica*, Vol. 93, No. 3: 959–1001

[39] Wilcox, Rand R., 2022, *Introduction to robust estimation and hypothesis testing*, 5th Edition, Academic Press.