



UNIVERSITAT
ROVIRA I VIRGILI

DEPARTAMENT D'ECONOMIA



WORKING PAPERS

Col·lecció “DOCUMENTS DE TREBALL DEL
DEPARTAMENT D'ECONOMIA - CREIP”

Judgement and Ranking: Living with Hidden Bias

António Osório

Document de treball n.24- 2016

DEPARTAMENT D'ECONOMIA – CREIP
Facultat d'Economia i Empresa



UNIVERSITAT
ROVIRA I VIRGILI

DEPARTAMENT D'ECONOMIA



Edita:

Departament d'Economia
www.fcee.urv.es/departaments/economia/public_html/index.html
Universitat Rovira i Virgili
Facultat d'Economia i Empresa
Av. de la Universitat, 1
43204 Reus
Tel.: +34 977 759 811
Fax: +34 977 758 907
Email: sde@urv.cat

CREIP
www.urv.cat/creip
Universitat Rovira i Virgili
Departament d'Economia
Av. de la Universitat, 1
43204 Reus
Tel.: +34 977 758 936
Email: creip@urv.cat

Adreçar comentaris al Departament d'Economia / CREIP

ISSN edició en paper: 1576 - 3382
ISSN edició electrònica: 1988 - 0820

DEPARTAMENT D'ECONOMIA – CREIP
Facultat d'Economia i Empresa

Judgement and Ranking: Living with Hidden Bias

António Osório[†]

[†]*Universitat Rovira i Virgili (Department of Economics) and CREIP
(antonio.osoriodacosta@urv.cat).*

Abstract

The complexity and subjectivity of the judgement task conceals the existence of biases that undermines the quality of the process. This paper presents a weighted aggregation function that attempts to reduce the influence of biased judgements on the final score. We also discuss a set of desirable properties. The proposed weighted aggregation function is able to correct the “nationalism bias” found by Emerson et al. (2009) in the 2000 Olympic Games diving competition and suggest the possibility of a “reputation bias”. Our results can be applied to judgement sports and other activities that require the aggregation of several personal evaluations.

Keywords: Weighted aggregation function; Judgement by Grading; Nationalism Bias; Reputation Bias; Bias Correction; Olympic Games.

JEL classification: D72, D81, D03.

1. Introduction

The functioning of our society frequently relies on evaluations and rankings about objects, performances or issues. These are made by qualified but potentially [conscious or unconscious] biased judges. Such is possible because the task is complex and inherently open to subjectivity and manipulation. The solution usually involves the aggregation of several evaluations into a single score.

The list of situations that require ranking based personal evaluations is endless. It includes wines, books, films and music contests, scientific refereeing or any kind of talent competition, which seem to be object of different sources of bias.¹ Nowadays,

¹These considerations are particular relevant in judgement sports as gymnastics, diving, skating, boxing, surfing or dressage, among others. In many websites users are asked to anonymously (or not) rate every kind of items, from touristic places to blogs comments.

internet is turning these evaluation procedures even more common. However, recent controversies have spurred researchers to explore bias in subjective judgments in more detail.²

Similarly, the list of behavioral biases is enormous. For instance, “nationalism bias” indicates a [usually conscious] tendency for judges to favor own country athletes (Emerson et al., 2009; Zitzewitz, 2006). On the other hand, “reputation bias” corresponds to a general [usually unconscious] tendency for judges to be influenced by the athlete’s reputation (Kingstrom and Mainstone, 1985; Findlay et al., 2004; among others).³ These bias are not merely an issue prevalent in subjective personal evaluations but inherent to every dimension of life (Buchanan et al., 1998). Tversky and Kahneman (1974) and Kahneman and Tversky (1996) describe them as shortcuts strategies to complex process information.

In addition to the limitations associated with subjective judgments, difficulties also arise at the aggregation stage. These issues dates back to Borda (1784), Condorcet (1785) and Laplace (1820), see Balinski and Laraki (2010). Range voting aggregation schemes are frequently used; voters rate each candidate with a grade within a specified interval. The candidate with the highest sum or average is the winner. The method has interesting properties and passes certain generalizations of the Arrow (1950) impossibility theorem but still open to strategic manipulation. This is actually its main limitation. Often a truncation is used to remove extreme scores and mitigate potential bias. In this respect, majority judgment (Balinski and Laraki, 2007; Balinski and Laraki, 2011) ranks candidates by the median score.⁴ They show that among the existing aggregation mechanisms it is the one that best resists manipulation and reduces the incentives to exaggerate. However, excessive truncation may leads to a loss of information and expressiveness that characterizes these score voting type schemes. For instance, in a typical population ranging from three to five appraisals, the removal of the highest and the lowest scores corresponds to an important loss of information. In particular, if bias is only a possibility.

²In this respect, some statistical based rating procedures have shown better results than expert opinions that are affected by several types of behavioral limitations (Dawes et al., 1989; Meehl, 1954). Other inconsistencies and paradoxical observations are reported in Ashenfelter and Quandt (1999), Fritz et al. (2012) and Hodgson (2008). A further development of these behavioral factors is behind the scope of the present paper.

³Other examples are “rank order bias” (Ginsburgh and Van Ours, 2003; among others) or “outlier aversion bias” in which judges avoid rating far from the mean of other judges (Lee, 2008).

⁴A particular case is Brams and Fishburn (1978) approval voting method in which only two different votes may be submitted.

The well-known [Gibbard \(1973\)](#) and [Satterthwaite \(1975\)](#) theorems point for the impossibility of designing one system that prevents all forms of bias.⁵ The present paper acknowledges this limitation, but simultaneously proposes a practical aggregation mechanism that attempts to reduce the bias influence on the final ranking.

In brief, our argument is the following. If a judge [conscious or unconscious] favors (respectively, penalizes) a particular candidate, she must be grading above (respectively, below) the mean score of the other judges. Therefore, the more (respectively, less) a score deviates from the others choices; lower (respectively, higher) must be its influence in the final score. We do not remove any particular score, instead, we use the information contained in the other judges grades to reduce the relevance of this particular score. Technically, the weight given to a particular judge is proportional to other judges' distance from the arithmetic mean grade. The formalization is done in [Section 2](#).

Our proposal is a refinement of the existent procedures. It does not dispense the complimentary and simultaneous use of transparency policies, as for instance, the public disclosure of each judge score, which are particularly powerful and simple anti-bias mechanisms. Note that even if judges have reputation concerns or exist outside monitoring, conscious bias does not disappear. Instead, it becomes more strategic and subtle, moving inside narrower intervals, making detection even more difficult. The aggregation mechanism in the present paper is a solution to these cases.

In [Section 3](#) we state and show a set of desirable properties. Finally, in [Section 4](#) we conclude with an illustrative application to the 2000 Olympic Games diving competition. We show that the weighted aggregation function is able to correct the “nationalism bias” found by [Emerson et al. \(2009\)](#). We also point the possibility that the athletes in question have simultaneously benefited from “reputation bias”.

2. The weighted aggregation function

In general there is no evidence to claim whether a particular judgement is biased or not. Therefore, we would like to have an aggregation mechanism that benefits from the expressiveness of the ranking type schemes, but simultaneously deals with the impossibility of strategy-proof-in-ranking and the existence of effective manipulation

⁵Another strand of the literature ([Miller et al., 2005](#); [Prelec, 2004](#); among others) proposes methods for eliciting truthful scoring in situations where objective truth is unknowable, see [Schlag et al. \(2015\)](#) for a survey. However, the operational applicability of these methods is limited by the prior informational requirements, assumptions and knowledge about individuals' true objective functions.

or bias. In other words, we want a mechanism that reduces the impact of extreme scores while minimizing the loss of information.

Another difficulty is the fact that judges' preferences are private information and impossible to determine. Therefore, the aggregation function must depend only on what can be known [in practice], which in many cases is very little.

With these objectives in mind, in what follows, we describe the aggregation mechanism suggested in the present manuscript.

Let $s_{i,j} \in [S_-, S_+] \subset \mathbb{R}$ be the grade or score award by judge $j \in J = \{1, \dots, n\}$ to competitor or candidate $i \in I = \{1, \dots, m\}$. Let $\mathbf{s}_i = (s_{i,1}, \dots, s_{i,n})$ denote the vector of grades award to individual $i \in I$. The competitor $i \in I$ arithmetic mean grade (over all judges $j \in J$) is denoted and defined as $\bar{s}_i \equiv \frac{1}{n} \sum_{j=1}^n s_{i,j}$. The weight given to the grade award by judge $j \in J$ to competitor $i \in I$ is a function, denoted as $w_{i,j} : [S_-, S_+]^n \rightarrow [0, 1]$, and defined as,

$$w_{i,j}(\mathbf{s}_i) \equiv \frac{\sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha}{(n-1) \sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha}, \quad (1)$$

where $\sum_{j=1}^n w_{i,j} = 1$, and $\alpha \geq 0$ controls for how much penalized are the grades that are far from the mean. In other words, the weight given to the grade of judge $j \in J$ depends on the grades of all other $l \neq j \in J$ judges and on their distances from the arithmetic mean.

Our objective is to penalize more the largest grade deviations from the mean, the ones that are most likely to be biased. However, a value of α that is too large can be problematic because bias is only a possibility and, in some cases, can be dissimulated away from the most extreme values. On the other hand, a low value of α may not penalize enough extreme values.⁶

Given a competitor $i \in I$, the Weighted Aggregation Function (WAF) weights by $w_{i,j}$ the grade $s_{i,j}$ of judge $j \in J$. We denote this function as $\bar{s}_i^w : [S_-, S_+]^n \rightarrow [S_-, S_+]$, and define it as,

$$\bar{s}_i^w(\mathbf{s}_i) \equiv \sum_{j=1}^n w_{i,j} s_{i,j}, \quad (2)$$

for all $i \in I$. In case of a tie between competitors the winner is the one with less dispersion around \bar{s}_i^w . Otherwise, ties are broken through a lottery.

⁶Our notion of distance does not correspond to a metric as it does not satisfy the triangle inequality, except for $\alpha \leq 1$.

Note that the particular cases $n = 1$ and $n = 2$ are consistent. In the latter, we always have $w_{i,1} = w_{i,2} = 1/2$, i.e., there is not enough information to make considerations regarding whether one judge is more or less correct than the other. The case $n = 1$ is trivial, i.e., $\bar{s}_i^w = s_{i,1}$. Another consistent particular case is when all grades coincide, in this case we have equal weighting, i.e., $w_{i,j} = 1/n$ for all $j \in J$, and $\bar{s}_i^w = \bar{s}_i = s_{i,j}$.⁷

In practice, it is natural to expect that the WAF correction mechanism is common knowledge among the judges. This knowledge allows consciously biased judges to strategically readjust their grades. However, in order to introduce bias in the WAF, the judges are forced to incur in more extreme behaviors, i.e., they must award grades that are more extreme than in the case in which the aggregation function is the arithmetic mean, for instance. This behavior exposes them to easier detection by third party monitoring. For that reason, the WAF in the present paper does not dispense public monitoring and the use of transparency policies as the public disclosure of the grade given by each judge.⁸

3. Properties

The proposed WAF satisfies the six axioms of the basic model, as defined by Balinski and Laraki (2007).⁹ Table 2 below resumes this information. In brief, *neutrality* and *anonymity* mean that any permutation of competitors and judges, respectively, does not affect the final grade. *Unanimity* means that if all judges award the same grade the final grade must be that grade. The symmetry property

⁷In this case, if $s_{i,j} = \bar{s}_i$ for all $j \in J$, we obtain an indetermination of the form $0/0$. Since the numerator and the denominator of $w_{i,j}$ are differentiable in some open interval around \bar{s}_i , equal weighting emerges after applying L'Hôpital's rule.

⁸The WAF defined in (2) with weights given by (1) is equivalent to the aggregation function $\bar{s}_i^w \equiv \sum_{j=1}^n w'_{i,j} \sum_{l \neq j}^n s_{i,l} / (n-1)$, with weights given by,

$$w'_{i,j} \equiv \frac{|s_{i,j} - \bar{s}_i|^\alpha}{\sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha}, \quad (3)$$

for all $i \in I$ and $j \in J$. Under this interpretation the intuition is reversed; the grade of a biased judge increases the relevance given to the grades of the other $n - 1$ non-biased judges. The reader is free to consider this alternative but equivalent formulation.

⁹Felsenthal and Machover (2008) provide a critical discussion on some additional properties of majority judgement and range voting, see also Balinski and Laraki (2014). We note that in some cases a direct transposition of some classical properties into ranking aggregation schemes is of interest but may ignore the contextual reality (Chebotarev and Shamis, 1998).

presented below is a more demanding generalization of *unanimity*, the implied equal treatment of every grade imply *neutrality* and *anonymity*. *Monotonicity* means that a higher grade cannot make the competitor in question worse-off. We will show that monotonicity holds in most cases. The exception occurs for large deviations from the mean and when these deviations are heavily punished, i.e., through a large value α . This property is discussed in more detail below. *Independence of irrelevant alternatives* means that the grades award to other competitors cannot affect directly the final grade of a given competitor. *Continuity* has the usual mathematical meaning. The proof of some of these properties is simple and follows from the definitions of weights in (1) and WAF in (2).

In this section, we state, prove and comment an additional set of desirable properties.

The aggregation function must weight equally grades that are at the same distance below and above the mean. Monitoring is achieved by considering deviations from the mean. Moreover, the WAF must return the arithmetic mean if there exist full symmetry between grades. A particular case is when all scores coincide.

Property 1 (symmetry). *If $s_{i,j} > \bar{s}_i$ and $s_{i,k} < \bar{s}_i$ such that $s_{i,j} - \bar{s}_i = \bar{s}_i - s_{i,k}$, then $w_{i,j} = w_{i,k}$. Moreover, if this is true for all $j, k \in J$, then $\bar{s}_i^w = \bar{s}_i$.*

This property implies an equal treatment for the grades that are below and above the mean. The satisfaction of this property is crucial because in general we cannot tell whether a particular grade is affected from bias, or whether this bias is below or above the mean. Moreover, a biased judge may benefit a particular competitor in several ways: directly through positive bias, indirectly through negative bias in the opponents of this competitor, or a combination of both. Therefore, we must be equally attentive to both sides of the mean.

The equal treatment of grades stated in the first part of Property 1 implies *neutrality* and *anonymity*. The second part of Property 1 implies *unanimity*. Note that these are three of the axioms of the basic model (Balinski and Laraki, 2007).

Once we have numerical grades we should be able to perform basic arithmetic transformations without changing their meaning. For instance, the distance between two points on a number line should not change by adding the same quantity to both numbers. The multiplication of two points on a number line does not have this property; distance is not invariant under multiplication. However, since the proposed weight function (1) is a ratio between sums of distances it is invariant under

multiplication. Consequently, the weights become invariant to linear operations in the grades.

Definition (scaling and translation invariant). *The weights are scaling and translation invariant if $w_{i,j}(\phi(\mathbf{s}_i)) = w_{i,j}(\mathbf{s}_i)$ for all $i \in I$ and $j \in J$, where $\phi(\mathbf{s}_i) = a + b\mathbf{s}_i$ with $a \in \mathbb{R}$, $b \in \mathbb{R}_+$ and $\mathbf{s}_i = (s_{i,1}, \dots, s_{i,n})$.*

Property 2 (scaling and translation invariant). *The weights are scaling and translation invariant.*

In other words, the weights are invariant to affine transformations of all grades. This property is passed to the WAF. Consequently, the WAF translates linearly changes in the grades.

Definition (scale-consistent). *The WAF is scale-consistent if $\bar{s}_i^w(\phi(\mathbf{s}_i)) = \phi(\bar{s}_i^w(\mathbf{s}_i))$ for all $i \in I$, where $\phi(\mathbf{s}_i) = a + b\mathbf{s}_i$ with $a \in \mathbb{R}$, $b \in \mathbb{R}_+$ and $\mathbf{s}_i = (s_{i,1}, \dots, s_{i,n})$.*

Property 3 (scale-consistent). *The WAF is scale-consistent.*

Scale-consistency is a less demanding property than language-consistency as it is defined in [Balinski and Laraki \(2010\)](#). The latter property refers not only to numerical grades but also to grades based on letters or descriptive phrases. This issue is related with the meaningfulness problem of measure theory in the context of jury decisions (see [Krantz et al. \(1971\)](#) for an early reference). In the present manuscript we focus on the aggregation of well-defined grades on numerical scales with no language-consistency issues among the judges.

A related but different property requires that the weight awarded by a given judge decreases with the distance of that grade from the mean.

Property 4 (decreasing weight). *Suppose that there exist at least two different grades other than $s_{i,j}$. Then $\partial w_{i,j} / \partial s_{i,j} > 0$ if $s_{i,j} < \bar{s}_i$ and $\partial w_{i,j} / \partial s_{i,j} < 0$ if $s_{i,j} \geq \bar{s}_i$ for all $i \in I$ and $j \in J$. Otherwise, $w_{i,j}$ is constant in $s_{i,j}$.*

The WAF decreases the impact of the most extreme values on the final grade by weighing them less than other grades that are closer to the mean. In our context, since the objective of the WAF is to correct for the possibility of bias this property is natural and the intuition is immediate. The more judge j moves away from the mean the less weight is given to her grade and more weight is given to the other

judges grades. This is the first statement of Property 4 and holds when there exist at least two different grades other than $s_{i,j}$.¹⁰

The second statement of Property 4 states that weights are constant when there is only one grade (or less) different from $s_{i,j}$.¹¹ In this case we may have two different scenarios:

- (i) In one scenario, we have $n - 1$ judges coinciding but judge j proposing something different. Then, it is natural to think that judge j is more likely to be biased than the other $n - 1$ judges because she is departing from the others. Consequently, the weight given to judge j becomes constant at some minimal value. The judge j is not ignored (because $w_{i,j}$ is strictly positive, see the proof of Property 4) but its importance is reduced to a minimum (judge j is only ignored if $\alpha \rightarrow \infty$, see Property 5 below). This property guarantees representativeness; some information may be less weighted but is not destroyed. In other words, since the weights are bounded from above and below, the WAF neither exclude any grade (i.e., $0 < w_{i,j}$) nor is based in a single grade (i.e., $w_{i,j} < 1$), with exception of the trivial case $n = 1$.
- (ii) In the other scenario, we have $n - 2$ judges coinciding with judge j but some judge k proposing something different. In this case the judge j is part of the group of judges that is more likely to be correct. Consequently, the grade of judge j receives a larger weight than in the previous scenario (but also constant).

However, the fact that $w_{i,j}$ might be constant in some particular cases does not imply that the WAF is constant. This issue is related with the monotonicity property discussed below.

In order to discussing monotonicity we consider two asymptotic properties of the WAF. These properties allow us to better understand the relation between the WAF and the value of α . Moreover, they provide a basis to understand why sometimes monotonicity fails.

Before we present these properties, we define *extreme grade* as the grade at the largest distance from the mean. Since extreme grades are not necessarily unique

¹⁰This statement could have been written without the resource to derivatives as follows: "(...) If $s_{i,j} > s_{i,k} > \bar{s}_i$, then $w_{i,j} > w_{i,k}$, while if $s_{i,j} < s_{i,k} \leq \bar{s}_i$, then $w_{i,j} < w_{i,k}$, for all $i \in I$ and $j, k \in J$." The reader may find this alternative but equivalent statement more intuitive.

¹¹The case in which all judges coincide is trivial; all weights are equal and the WAF delivers the common grade.

and equal, we also define the *average extreme grade* as the arithmetic mean over all extreme grades.

Property 5 (asymptotic cases). *If $\alpha \rightarrow 0$ then $\bar{s}_i^w \rightarrow \bar{s}_i$. If $\alpha \rightarrow \infty$ then $\bar{s}_i^w \rightarrow (\sum_{l=1}^n s_{i,l} - \bar{s}_{i,e})/(n-1)$, where $\bar{s}_{i,e}$ denotes the average extreme grade.*

In the limit case $\alpha \rightarrow 0$, the WAF converges to the mean grade. The WAF weights all grades equally independently of their distance from the mean (i.e., $w_{i,j} \rightarrow 1/n$ for all $j \in J$).

In the limit case $\alpha \rightarrow \infty$, the WAF ignores the influence of the average extreme grade. Intuitively, if there is only one extreme grade, the WAF ignores this grade - the one that is more likely to be biased - and averages with equal weight the other $n-1$ grades. In the most general case, we may consider the possibility of several extreme grades. If there are r extreme grades, the WAF weights each of these r extreme grades by $(r-1)/((n-1)r)$ and each of the other $n-r$ non-extreme grades by $1/(n-1)$. The intuition is the same as in the $r=1$ case, but we have two different situations. For instance, in the profile of grades (6, 6, 7, 9, 9) since $\bar{s}_i = 7.4$, we have two equal extreme grades - both in the same side of the mean, i.e., $s_{i,4} = s_{i,5} = 9$. In the profile of grades (5, 6, 7, 8, 9) since $\bar{s}_i = 7$, we have two but unequal extreme grades - one in each side of the mean, i.e., $s_{i,1} = 5$ and $s_{i,5} = 9$.

These observations imply that the WAF cannot be monotonic in a single judge grade for all values of α . In order to see it, consider the following example.

Example 1. *Consider the profile of grades (5, 6, 6.9) with $\bar{s}_i = 5.97$. Suppose that α is very large (say $\alpha \approx \infty$) such that the most extreme grade is removed from \bar{s}_i^w according to Property 5. In this case, the grade $s_{i,1} = 5$ has the largest distance from the mean and for that reason is removed from $\bar{s}_i^w \rightarrow (6 + 6.9)/2 = 6.45$. However, if $s_{i,3}$ increases from 6.9 to 7.1, then $s_{i,3}$ becomes the most extreme grade and for that reason is removed from $\bar{s}_i^w \rightarrow (5 + 6)/2 = 5.5$. In this example, after the increase of $s_{i,3}$, the WAF falls abruptly from $\bar{s}_i^w = 6.45$ to $\bar{s}_i^w = 5.5$ which shows the failure of monotonicity with respect to $s_{i,3}$.*

The example was constructed specially to show the failure of monotonicity in a simple way and to motivate the discussion about the meaning and importance of this property in a bias correction mechanism. This is a crucial aspect; in some sense we allow a judge to have full control over her own grade but we only allow partial and restricted influence over the final grade. In this context, we do not exclude the possibility that the bias correction mechanism reacts to extreme grades with a

detrimental effect on the final score. In other words, if a judge awards a competitor with a grade that is too large (respectively, small), with respect to the other judges, the final grade may decrease (respectively, increase). Consequently, the existence of a monotonic relation between a particular grade and the final score is not always guaranteed.

The following property states under which circumstances the WAF is monotonic.

Property 6 (monotonicity). \bar{s}_i^w is monotonic in $s_{i,j}$ if α is sufficiently small.
 \bar{s}_i^w is monotonic in $s_{i,j}$ if $s_{i,j}$ is sufficiently close to the other judges mean.

The numerical example in Table 1 provides an illustration of Property 6 for $\alpha = 3$. Column (6) of Table 1 shows that while the grade of the first judge raises from $s_{i,1} = 6$ to $s_{i,1} = 7$ or from $s_{i,1} = 7$ to $s_{i,1} = 8$ we observe a monotonic relation between $s_{i,1}$ and \bar{s}_i^w . In this case, $s_{i,1}$ remains sufficiently close to the mean. However, when the grade of the first judge raises from $s_{i,1} = 8$ to $s_{i,1} = 9$ the WAF falls from $\bar{s}_i^w = 7.314$ to $\bar{s}_i^w = 7.303$ and monotonicity fails. This movement is due to the fall in the weight given to the first judge relatively to the increase in the grade (the weight decreases from $w_{i,1} = 0.204$ to $w_{i,1} = 0.107$, see Column (1) of Table 1).

The failure of monotonicity occurs because the WAF reacts to extreme grades (the ones that more likely to have been object of bias) through a correction in the opposite direction. In such case an extreme positive (respectively, negative) grade reduces (respectively, rises) the final grade instead of increase (respectively, decrease) it. In our context, since the objective is to propose a bias correction mechanism the failure of monotonicity can be seen as a desirable property. This observation gains even more strength if we consider the circumstances under which it occurs. In particular, monotonicity fails if α is large enough or $s_{i,j}$ is sufficiently distant from the mean.

Note that the choice of α reflects an explicit intention of punishing deviations from the mean. In our context, this parameter is controlled by the social planner or the competition designer. The larger the value of α , the stronger is the reduction in the weight given to the grades that are distant from the mean. However, if we are more tolerant regarding the grades that are distant from the mean the monotonic relation between \bar{s}_i^w and $s_{i,j}$ is likely to be satisfied.

Note also that a necessary condition for monotonicity to fail is that $s_{i,j}$ is an extreme grade. Otherwise, inside the interval bounded by the largest and the lowest grade, the monotonic relation between \bar{s}_i^w and $s_{i,j}$ is satisfied. In some sense, the WAF has a self-correction mechanism against extreme grades. Column (6) of Table

1 shows that monotonicity fails when the first judge raises her grade above the other judges largest grade which in this case is $s_{i,5} = 8$.

In spite of the failure of monotonicity being convenient, because it occurs under extreme circumstances, there are also negative aspects associated with it. For instance, if a given judge ranks individual A above B then the bias correction mechanism will not necessarily respect this ordering. Consequently, if that judge is removed from the average then individual A can improve over individual B . This effect is known in the literature as the *no-show paradox* (Felsenthal and Machover, 2008; Fishburn and Brams, 1983; Smith, 1973).

profile	$w_{i,1}$ (1)	$w_{i,2}$ (2)	$w_{i,3}$ (3)	$w_{i,4}$ (4)	$w_{i,5}$ (5)	\bar{s}_i (6)	\bar{s}_i^w (7)	skew
(6,6,7,7,8)	0.204	0.204	0.249	0.249	0.094	6.800	6.686	+
(7,6,7,7,8)	0.250	0.125	0.250	0.250	0.125	7.000	7.000	0
(8,6,7,7,8)	0.204	0.094	0.249	0.249	0.204	7.200	7.314	-
(9,6,7,7,8)	0.107	0.155	0.248	0.248	0.242	7.400	7.303	+

Table 1: **Weights, WAF, Mean, Median and Skewness** (case $\alpha = 3$): The effect of variations in the first judge grade when the grades of the other four judges is constant.

Now, we consider another property of the WAF that is also related with Property 5.

Recall that the proposed aggregation function has the objective of controlling for individual bias. In this context, the correction mechanism gives more weight to the opinion of the judges that are assumed not to be biased. The question is, which are the most likely unbiased judges? Without further information, the answer points to the judges whose grades show higher prevalence and similitude. This idea has motivated the WAF proposed in the present paper. An implication is the following property.

Property 7 (homogeneity stickiness). *If the distribution of grades has positive (negatively, respectively) skewness then $\bar{s}_i^w < \bar{s}_i$ ($\bar{s}_i^w > \bar{s}_i$, respectively).*

The WAF compensates positive skewness by lowering \bar{s}_i^w below \bar{s}_i , and does the opposite otherwise. This property expresses a movement of the final grade towards the most similar and frequent grades of the distribution. The most homogeneous majority has a greater decisiveness in the final grade. This property is important

because a large group of individuals with more even grades is more likely to be correct and more difficult to manipulate than a small group.

As a rule of thumb positive skewness implies $median < \bar{s}_i$ and negative skewness implies $\bar{s}_i < median$. However, this rule sometimes fails (e.g., for multimodal distributions or distributions where one tail is long but the other is heavy). In our case, this rule fails for discrete distributions in which the areas to the left and the right of the median are not equal. This observation highlights situations in which the WAF departure from the median is immediate because in those cases the median does not adjust in the correct direction. In order to better understand this point, consider the numerical example in Table 1. Part 1 highlights the difference between the median and the WAF. Part 2 shows how the homogeneity stickiness Property 7 works and establishes the connection with Property 5.

Example 2. (Part 1) Suppose that $n = 5$ and $\mathbf{s}_i = (6, 6, 7, 7, 8)$. In this case the distribution of grades is positively skewed (Last column in Table 1). The WAF gives more weight to the most homogenous set of grades $(6, 6, 7, 7, .)$ (Columns (1)-(5) of Table 1) and we have $\bar{s}_i^w = 6.686 < \bar{s}_i = 6.800$ (Columns (6) and (7) of Table 1). However, the median moves in the opposite direction, i.e., $\bar{s}_i = 6.800 < median = 7$.

(Part 2) In the continuation of Part 1, if the first judge awards $s_{i,1} = 8$, the distribution of grades becomes negatively skewed. The WAF gives more weight to the largest grades $(8, ., 7, 7, 8)$ because they form the most homogeneous block. However, if the first judge awards $s_{i,1} = 9$, the distribution of grades becomes positively skewed again. The WAF gives more weight to the lowest four grades $(., 6, 7, 7, 8)$. Finally, note that according to Property 5, if $\alpha \rightarrow \infty$ then \bar{s}_i^w is the mean of the most homogeneous block of four grades.

This example is not stylized. With five judges the profiles of grades presented in Example 2 are extremely common in judgement sports as gymnastics, diving, skating, boxing, surfing or dressage, among others.

Finally, Properties 1-7 cannot uniquely characterize the proposed aggregation method, mostly because in our setting the weights are not constant but depend in a non-trivial way on the grades that they are weighting, see expressions (1) and (2). [Beliakov et al. \(2007\)](#) and [Grabisch et al. \(2011a,b\)](#) survey this literature. However, in general Properties 1-7 are not easy to satisfy by other aggregation methods because of nonlinearities and cross effects.

	Weighted Aggregation Function
Neutrality	Yes
Anonymity	Yes
Unanimity	Yes
Monotonicity	Yes ^{b)}
Independent of irrelevant alternatives	Yes
Continuity	Yes
Symmetry	Yes
Scaling and translation invariance ^{a)}	Yes
Scale-consistency	Yes
Decreasing weight ^{a)}	Yes
Homogeneity stickiness	Yes

Table 2: **General properties of the *Weighted Aggregation Function*:** ^{a)} these properties refer to the weights, ^{b)} this property is true under certain conditions.

4. An Application

The diver Fernando Platas of Mexico lost the 2000 Olympic gold medal in the 3-meter springboard diving competition with an extremely narrow margin to Xiong Ni of China (Column (1) of Table 3). The result generated controversy among fans and press because from the eleven dives counting for the final ranking, three were graded by the Chinese judge Facheng Wang during the semi-final stage.¹² Some years later, Emerson et al. (2009) could not statistically reject the hypothesis that Xiong Ni has benefited from nationalistic judging bias.

The method used by the International Olympic Committee to compute the final score was the following one. The judging panel was composed by seven judges. Each dive final grade is calculated by summing the middle five awards (the lowest and highest scores were removed) and then multiplying the obtained number by the degree of difficulty DD_i and by $3/5$, according to the following formula:

$$points_i = DD_i \times \frac{3}{5} \times \left(\sum_{j=1}^7 s_{i,j} - \min_j \{s_{i,j}\} - \max_j \{s_{i,j}\} \right),$$

for all $i \in I$.

¹²Judges from competitors' countries were not assigned to the final - but they could be in earlier rounds, as in this case, in the semi-final stage.

Our results corroborate [Emerson et al. \(2009\)](#). The application of the WAF (with $\alpha = 2$) to the grades award in the eleven dives returns that Fernando Platas would have won the gold medal with 709.61 points against the Ni Xiong 709.39 points (Column (2) in Table 3).

	Rank	Points (1)	WAF (2)	MJ (3)	(4) = (2)-(1)
Xiong Ni	1	708.72	709.39	713.25	0.67
Fernando Platas	2	708.42	709.61	712.20	1.19
Dmitri Sautin	3	703.02	704.29	699.90	1.09
Xiao Hailiang	4	671.04	671.06	674.10	0.02
Dean Pullar	5	647.40	647.94	645.30	0.54
Troy Dumais	6	642.72	641.62	641.85	-1.10
Mark Ruiz	7	638.22	636.84	631.50	-1.38
Ken Terauchi	8	634.47	633.26	630.60	-1.22
Stefan Ahrens	9	619.17	616.94	617.10	-2.23
Andreas Wels	10	616.53	613.92	616.20	-2.61
Imre Lengyel	11	613.47	613.67	616.20	0.20
Tony Ally	12	583.80	585.00	588.60	1.20

Table 3: **The Olympic Committee versus the Weighted Aggregation Function.** Source: Emerson et al. (2009).

Note that the International Olympic Committee and the WAF scoring rules are very similar (Columns (1) and (2) of Table 3, respectively). This aspect is important because we want the WAF to correct for potential bias but not to affect indiscriminately the results because bias is only a possibility. In some sense our proposal is a refinement of the International Olympic Committee procedure, and does not disperse transparency and public disclosure policies which are particularly simple and powerful anti-bias monitoring mechanisms.

Note that majority judgement (Column (3) of Table 3) preserves the International Olympic Committee ranking. Therefore, failing to capture the nationalistic judging bias in favor of the diver Xiong Ni found by [Emerson et al. \(2009\)](#). Regarding this aspect, majority judgement is a particularly powerful aggregation method against bias and manipulation. The failure in this particular case is due to the asymmetry of the median method; see Part 1 of Example 2 above. Note that when we remove one grade from the left and one grade from the right of the median these two grades can

be very asymmetric. In the case of great asymmetry there is a loss of information and in this particular case explains why majority judgement was not able to remove the existent bias. The situation considered in Table 3 is just a particular example; in other situations majority judgement may effectively remove bias.

Finally, the comparison of the difference between the WAF and the International Olympic Committee scores (Column (4) of Table 3) suggests the possibility of another different source of bias.¹³ The divers ranked from the first to the fifth position present a positive difference while the following five divers show a negative difference. This pattern seems to suggest a tendency for the judges to benefit the better positioned divers - reputation bias. If this was the case, both athletes (Fernando Platas and Ni Xiong) have been benefited in this dimension. The joint consideration of these two forms of bias (i.e., nationalistic and reputation bias) may return the first place back to Xiong Ni. The statistical validation of this claim requires information that is not easily available. Nonetheless, further research on these issues is of great relevance to understand the behavioral forces that shape judges decisions. The study of multiple and simultaneous sources of bias seems to be a particularly interesting and unexplored subject.

Acknowledgments: I wish to thank to Ricardo Ribeiro, Juan Pablo Rincón-Zapatero, the Associate Editor and two anonymous referees, as well as several seminars and congresses participants for helpful comments and discussions. Financial support from the Spanish Ministerio of Ciencia y Innovación, GRODE and the Barcelona GSE is gratefully acknowledged. The usual caveat applies.

¹³It is interesting to note that the aggregation function may have a dual interpretation. If $\bar{s}_i^w < \bar{s}_i$ (respectively, $\bar{s}_i^w > \bar{s}_i$) the mechanism corrects for the possibility that candidate i is benefited (respectively, penalized) by a single judge bias. However, we can reverse the argument and assume that the minority is correct and look for the existence of a majority bias. In this case, $\bar{s}_i^w < \bar{s}_i$ (respectively, $\bar{s}_i^w > \bar{s}_i$) suggests that candidate i might have been penalized (respectively, benefited) by a generalized bias (e.g., “reputation bias” - judges are influenced by the athlete’s reputation). This feature of the WAF is particularly interesting. However, the WAF defined in (2) with weights as in (1) does not perform the correction. Such correction could be done by replacing the weight function (3) into (2). In this case, the hypothesis is that there are $n - 1$ biased judges. Therefore, their grades should lose relevance relatively to the grade of the unbiased judge $j \in J$. Similar reasoning can be applied to any other number of suspected biased judges.

Appendix

Proof of Property 1. In order to show the first statement just note that if $s_{i,j} - \bar{s}_i = \bar{s}_i - s_{i,k}$ by the definition of absolute value function we must have $|s_{i,j} - \bar{s}_i|^\alpha = |s_{i,k} - \bar{s}_i|^\alpha$. Since all other terms in $w_{i,j}$ and $w_{i,k}$ are similar, we must have $w_{i,j} = w_{i,k}$. The second statement implies that if for every $s_{i,j}$ exist another $s_{i,k}$ at the same distance from the mean, say $s_{i,j} - \bar{s}_i = \bar{s}_i - s_{i,n+1-j}$ which can be rewritten as $s_{i,j} + s_{i,n+1-j} = 2\bar{s}_i$, then by the first statement $w_{i,j} = w_{i,n+1-j}$, and we have:

$$\bar{s}_i^w = \sum_{j=1}^n w_{i,j} s_{i,j} = \sum_{j=1}^{n/2} w_{i,j} (s_{i,j} + s_{i,n+1-j}) = \sum_{j=1}^{n/2} 2w_{i,j} \bar{s}_i = \bar{s}_i,$$

because $\sum_{j=1}^{n/2} w_{i,j} = 1/2$. In the case that the total number of grades is odd, then there must exist a odd number of grades equal to the mean in order to exist symmetry. ■

Proof of Property 2. In order to show scaling and translation invariance, apply the affine transformation function $\phi(s_{i,j}) = a + bs_{i,j}$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}_+$, to each grade $s_{i,j}$ in the weight function $w_{i,j}$, given by expression (1), to obtain:

$$\begin{aligned} w_{i,j}(\phi(\mathbf{s}_i)) &= \frac{\sum_{l \neq j}^n |a + bs_{i,l} - \frac{1}{n} \sum_{k=1}^n (a + bs_{i,k})|^\alpha}{(n-1) \sum_{l=1}^n |a + bs_{i,l} - \frac{1}{n} \sum_{k=1}^n (a + bs_{i,k})|^\alpha} \\ &= \frac{\sum_{l \neq j}^n |a + bs_{i,l} - a - \frac{1}{n} b \sum_{k=1}^n s_{i,k}|^\alpha}{(n-1) \sum_{l=1}^n |a + bs_{i,l} - a - \frac{1}{n} b \sum_{k=1}^n s_{i,k}|^\alpha} \\ &= \frac{\sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha}{(n-1) \sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha} = w_{i,j}(\mathbf{s}_i), \end{aligned}$$

for all $i \in I$ and $j \in J$. ■

Proof of Property 3. In order to show scale-consistency, apply the affine transformation function $\phi(s_{i,j}) = a + bs_{i,j}$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}_+$, to each grade $s_{i,j}$ in the WAF $\bar{s}_i^w(\mathbf{s}_i)$, given in expression (2). Since by Property 2 weights are scaling and translation invariant it is enough to show that:

$$\bar{s}_i^w(\phi(\mathbf{s}_i)) = \sum_{j=1}^n w_{i,j} (a + bs_{i,j}) = a + b \sum_{k=1}^n w_{i,j} s_{i,j} = \phi(\bar{s}_i^w(\mathbf{s}_i)),$$

because $\sum_{j=1}^n w_{i,j} = 1$ for all $i \in I$. ■

Proof of Property 4. We start by considering the second statement of Prop-

erty 4, i.e., the case in which there is only one grade different from $s_{i,j}$.¹⁴ We have two different scenarios. (i) In the first scenario, $n - 1$ judges coincide but judge j proposes something different. Let $s_{i,k} = s_{i,m}$ for all $k \neq j \in J$, then, we can write $\sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha = (n - 1) |s_{i,m} - \bar{s}_i|^\alpha$. After having replaced it in expression (1) we obtain:

$$w_{i,j} = \frac{(n - 1) |s_{i,m} - \bar{s}_i|^\alpha}{(n - 1)(|s_{i,j} - \bar{s}_i|^\alpha + (n - 1) |s_{i,m} - \bar{s}_i|^\alpha)} = \frac{1}{(n - 1)^\alpha + (n - 1)} \equiv w^{\min}, \quad (4)$$

where in the last equality we made use of the fact that in this case $|s_{i,j} - \bar{s}_i| = (n - 1)|s_{i,m} - \bar{s}_i|$ for $n > 1$. Consequently, the weight given to judge j is constant at some minimal value w^{\min} . (ii) In the second scenario, $n - 2$ judges coincide with judge j but some other judge k proposes something different. Consequently, judge j grade receives a larger weight than w^{\min} because she is in the group of judges that is likely to be correct, i.e., we have $w_{i,j} = (1 - w_{i,k})/(n - 1) = (1 - w^{\min})w^{\max}$.¹⁵ In both cases $w_{i,j}$ is constant with $s_{i,j}$.

Now, we consider the case in which there are at least two grades different from $s_{i,j}$, i.e., the first statement of Property 4. In this case $w_{i,j}$ is given by expression (1). In order to reduce the size of the expression of the derivative $\partial w_{i,j}/\partial s_{i,j}$, let $z_{i,j} = \text{sgn}[s_{i,j} - \bar{s}_i] \in \{-1, 0, 1\}$ denote the sign function where $s_{i,j} > \bar{s}_i$ implies $z_{i,j} = 1$, $s_{i,j} = \bar{s}_i$ implies $z_{i,j} = 0$ while $s_{i,j} < \bar{s}_i$ implies $z_{i,j} = -1$. Then, differentiate $w_{i,j}$ with respect to $s_{i,j}$ to obtain:

$$\frac{\partial w_{i,j}}{\partial s_{i,j}} = \frac{-\alpha \frac{1}{n} |s_{i,j} - \bar{s}_i|^\alpha \sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1} - \alpha \left(1 - \frac{1}{n}\right) z_{i,j} |s_{i,j} - \bar{s}_i|^{\alpha-1} \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha}{(n - 1) \left(\sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha\right)^2}, \quad (5)$$

for all $i \in I$ and $j \in J$, where we made use of the fact that $\sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha = |s_{i,j} - \bar{s}_i|^\alpha + \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha$ and that the derivative of $\sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha$ with respect to $s_{i,j}$ is given by $\alpha z_{i,j} (1 - 1/n) |s_{i,j} - \bar{s}_i|^{\alpha-1} - \alpha (1/n) \sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1}$. We want to show that $\partial w_{i,j}/\partial s_{i,j} > 0$ if $s_{i,j} < \bar{s}_i$ ($z_{i,j} < 0$) and $\partial w_{i,j}/\partial s_{i,j} < 0$ otherwise. Since the denominator is strictly positive the sign of the derivative is given by the numerator. We have four cases to consider that depend on whether $z_{i,j}$ and $\sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1}$

¹⁴The case in which all judges coincide is trivial because all weights are equal.

¹⁵We can show that the maximum weight $w^{\max} = 1/(n - 1)$, for $n > 2$, is given to the judge or judges awarding grades that match the arithmetic mean. If all judges award equal grades they are all equally weighted. If $n = 2$ we always have $w_{i,j} = w_{i,k} = 1/2$ because both judge are equally likely to be correct.

are negative or positive. If $z_{i,j} > 0$ and $\sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1} > 0$, it is immediate that $\partial w_{i,j} / \partial s_{i,j} < 0$, while if $z_{i,j} < 0$ and $\sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1} < 0$, it is immediate that $\partial w_{i,j} / \partial s_{i,j} > 0$. In the other two cases, if $z_{i,j} > 0$ and $\sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1} < 0$, in order to show that $\partial w_{i,j} / \partial s_{i,j} < 0$ we must consider the scenario that makes it more difficult to satisfy, i.e., when $z_{i,l} = -1$ for all $l \neq j$. Similarly, if $z_{i,j} < 0$ and $\sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1} > 0$, in order to show that $\partial w_{i,j} / \partial s_{i,j} > 0$ we must consider the scenario that makes it more difficult to satisfy, i.e., when $z_{i,l} = 1$ for all $l \neq j$. In both cases we are left to show that the same inequality is true:

$$(n-1) \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha > |s_{i,j} - \bar{s}_i| \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^{\alpha-1},$$

for all $i \in I$ and $j \in J$. Note that for the case $s_{i,j} = \bar{s}_i$ the inequality holds trivially. Since the WAF is scale-consistent (Property 3) we can normalize $\bar{s}_i = 0$ without loss of generality and for simplicity consider the case of two different grades other than $s_{i,j}$, i.e., $s_{i,j} < \bar{s}_i = 0 < s_{i,k}, s_{i,l}$ (the other case $s_{i,j} > \bar{s}_i = 0 > s_{i,k}, s_{i,l}$ follows the same argument). Then, the above inequality becomes:

$$2(|s_{i,k}|^\alpha + |s_{i,l}|^\alpha) > (|s_{i,k} + s_{i,l}|)(|s_{i,k}|^{\alpha-1} + |s_{i,l}|^{\alpha-1}),$$

where we have used the fact that $\bar{s}_i = 0$ implies that $-s_{i,j} = s_{i,k} + s_{i,l}$. Since all quantities are non-negative we can remove the absolute value function from the previous inequality to obtain $(s_{i,k}^{\alpha-1} - s_{i,l}^{\alpha-1})(s_{i,k} - s_{i,l}) > 0$ which is strictly positive for all $s_{i,k} \neq s_{i,l}$ with $k \neq l \in J$. ■

Proof of Property 5. We start by considering the limit $\alpha \rightarrow 0$. Suppose that there are $t = 0, 1, \dots, n-1$ grades such that $s_{i,k} = \bar{s}_i$, denoted with the subindex $k \in K \subset J$ (the case $t = n$ is trivially true that $\bar{s}_i^w = \bar{s}_i$), and $n-t$ grades such that $s_{i,j} \neq \bar{s}_i$, denoted with the subindex $j \in J \setminus K$. Then, if $\alpha \rightarrow 0$ we have $|s_{i,j} - \bar{s}_i|^\alpha \rightarrow 1$ for all $j \in J \setminus K$, and $|s_{i,k} - \bar{s}_i|^\alpha \rightarrow 0$ for all $k \in K$. Consequently, $\sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha \rightarrow n-1-t$, $\sum_{l \neq k}^n |s_{i,l} - \bar{s}_i|^\alpha \rightarrow n-t$ and $\sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha \rightarrow n-t$, which implies that $w_{i,j} \rightarrow (n-1-t)/((n-t)(n-1))$ for all $j \in J \setminus K$ and $w_{i,k} \rightarrow 1/(n-1)$ for all $k \in K$. Therefore, we can write $\bar{s}_i^w \rightarrow \sum_{j \notin K} (n-1-t)s_{i,j}/((n-t)(n-1)) + ts_{i,k}/(n-1)$, where $s_{i,k} = \bar{s}_i = \sum_{j=1}^n s_{i,j}/n$, which implies that $s_{i,k} = \bar{s}_i = \sum_{j \notin K} s_{i,j}/(n-t)$. Replacing the latter two equalities into \bar{s}_i^w , after some algebra, we obtain that $\bar{s}_i^w \rightarrow \bar{s}_i$.

Now, consider the limit $\alpha \rightarrow \infty$. In this case we can rewrite $w_{i,j}$ as $w_{i,j} = 1/((n-1)(1 + |s_{i,j} - \bar{s}_i|^\alpha / \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha))$, for all $j \in J$. Suppose that there are $r = 1, \dots, \lfloor (n-1)/2 \rfloor$ extreme grades (not necessarily equal but at the same distance from the mean), where $\lfloor x \rfloor$ denotes floor function - the largest integer less than or equal to x (note: we cannot have more than $\lfloor (n-1)/2 \rfloor$ extreme grades, otherwise,

they will not be the extreme grades). There are two different situations to consider. (i) If $|s_{i,j} - \bar{s}_i|$ is one of the largest grade differences with respect to the mean (i.e., $s_{i,j}$ is an extreme grade) and $\alpha \rightarrow \infty$, then $|s_{i,j} - \bar{s}_i|^\alpha / \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha \rightarrow 1/(r-1)$ and

$$w_{i,j} \rightarrow 1/((n-1)(1 + 1/(r-1))) = (r-1)/((n-1)r).$$

Note that in the particular case $r = 1$ we have $w_{i,j} \rightarrow 0$. (ii) If $|s_{i,j} - \bar{s}_i|$ is not one of the largest grade differences with respect to the mean (i.e., $s_{i,j}$ is not an extreme grade) and $\alpha \rightarrow \infty$, then $|s_{i,j} - \bar{s}_i|^\alpha / \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha \rightarrow 0$ and $w_{i,j} \rightarrow 1/(n-1)$ for all $j \neq k \in J$. Altogether, we have $n-r$ non-extreme grades, each weighted by $1/(n-1)$ and r extreme grades, each weighted by $(r-1)/((n-1)r)$. Let the r extreme grades be indexed as $j = n-r+1, \dots, n$, then $\bar{s}_i^w \rightarrow (\sum_{j=1}^{n-r} s_{i,j} + (r-1) \sum_{j=n-r+1}^n s_{i,j}/r)/(n-1)$. After adding and subtracting $\sum_{j=n-r+1}^n s_{i,j}$ in the numerator we obtain that $\bar{s}_i^w \rightarrow (\sum_{j=1}^n s_{i,j} - \sum_{j=n-r+1}^n s_{i,j}/r)/(n-1)$, where $\sum_{j=n-r+1}^n s_{i,j}/r$ is the average extreme grade $\bar{s}_{i,e}$. ■

Proof of Property 6. We start by showing the first statement of Property 6, i.e., \bar{s}_i^w is monotonic in $s_{i,j}$ for small α . In order to do it we proceed as follows. Since $\alpha \in [0, \infty)$ we will show that \bar{s}_i^w is monotonic increasing in $s_{i,j}$ for α small, i.e., in the zero neighborhood. Then, through a numerical example, we show that for sufficiently large α the monotonic relation between \bar{s}_i^w and $s_{i,j}$ is not guaranteed. Note that the WAF defined in (2) can be written as (see Footnote 8):

$$\bar{s}_i^w = \left(\sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha \sum_{k \neq l}^n s_{i,k} \right) / \left((n-1) \sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha \right) = N/D,$$

for all $i \in I$ and $j \in J$, where N and D denote the expressions in the numerator and denominator, respectively. Recall that the derivative of \bar{s}_i^w with respect to $s_{i,j}$ is given by $\partial \bar{s}_i^w / \partial s_{i,j} = (N'D - ND')/D^2$. In the case that $s_{i,j} \neq \bar{s}_i$ for all $j \in J$ (the case that $s_{i,k} = \bar{s}_i$ for some $k \in J$ follows the same argument) the derivative of the expression in the denominator with respect to $s_{i,j}$ is equal to the numerator of expression (5) multiplied by $n-1$, that is:

$$D' = (n-1)\alpha z_{i,j}(1 - 1/n) |s_{i,j} - \bar{s}_i|^{\alpha-1} - (n-1)\alpha(1/n) \sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1},$$

where we use the same notation as in the Proof of Property 4. While the expression in the numerator can be written as $N = |s_{i,j} - \bar{s}_i|^\alpha \sum_{k \neq j}^n s_{i,k} + \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha \sum_{k \neq l}^n s_{i,k}$.

The derivative of this expression with respect to $s_{i,j}$ is given by:

$$\begin{aligned} N' &= \alpha z_{i,j}(1 - 1/n) |s_{i,j} - \bar{s}_i|^{\alpha-1} \sum_{k \neq j}^n s_{i,k} \\ &\quad - \alpha(1/n) \sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1} \sum_{k \neq l}^n s_{i,k} + \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha. \end{aligned}$$

In the limit $\alpha \rightarrow 0$, we have $N \rightarrow (n-1) \sum_{l=1}^n s_{i,l}$, $D \rightarrow (n-1)n$, $N' \rightarrow n-1$ and $D' \rightarrow 0$. Therefore, $\partial \bar{s}_i^w / \partial s_{i,j} \rightarrow 1/n$ for $\alpha \rightarrow 0$ which is strictly positive for any profile of grades. For large values of α we show through a numerical example the failure of monotonicity. Column (6) of Table 1 shows that for $\alpha = 3$ when the grade of the first judge raises from $s_{i,1} = 8$ to $s_{i,1} = 9$ the WAF falls from $\bar{s}_i^w = 7.314$ to $\bar{s}_i^w = 7.303$, showing that monotonicity is not guaranteed for large α .

Now, we consider the second statement of Property 6. The proof of monotonicity for $s_{i,j}$ around the mean follows a similar strategy. We evaluate $\partial \bar{s}_i^w / \partial s_{i,j}$ in the neighborhood of \bar{s}_i , i.e., for $s_{i,j} \rightarrow \bar{s}_i$ which is equivalent to $s_{i,j} \rightarrow \sum_{l \neq j}^n s_{i,l} / (n-1)$ (the other judges mean). In this case we have $|s_{i,j} - \bar{s}_i|^\alpha \rightarrow 0$, and consequently, $N \rightarrow \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha \sum_{k \neq l}^n s_{i,k}$, $D \rightarrow (n-1) \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha$,

$$N' \rightarrow -\alpha(1/n) \sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1} \sum_{k \neq l}^n s_{i,k} + \sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha,$$

and $D' \rightarrow -(n-1)\alpha(1/n) \sum_{l \neq j}^n z_{i,l} |s_{i,l} - \bar{s}_i|^{\alpha-1}$. At this stage the expression for $\partial \bar{s}_i^w / \partial s_{i,j}$ is particularly large. Since the WAF is scale-consistent (Property 3) we can normalize $\bar{s}_i = 0$ without loss of generality, which implies that $s_{i,j} \rightarrow 0$ and we can write $\sum_{k \neq l}^n s_{i,k} = -s_{i,l}$. Then, the expression for $\partial \bar{s}_i^w / \partial s_{i,j}$ converges to:

$$\frac{(-\alpha \frac{1}{n} \sum_{l \neq j}^n z_{i,l} |s_{i,l}|^{\alpha-1} (-s_{i,l}) + \sum_{l \neq j}^n |s_{i,l}|^\alpha \sum_{l \neq j}^n |s_{i,l}|^\alpha + \alpha \frac{1}{n} \sum_{l \neq j}^n |s_{i,l}|^\alpha (-s_{i,l}) \sum_{l \neq j}^n z_{i,l} |s_{i,l}|^{\alpha-1})}{(n-1)(\sum_{l \neq j}^n |s_{i,l}|^\alpha)^2}, \quad (6)$$

for all $i \in I$ and $j \in J$. Note that $s_{i,l} > 0$ is equivalent to $z_{i,l} > 0$ (positive and above the mean) and implies that $z_{i,l} |s_{i,l}|^{\alpha-1} (-s_{i,l}) = -|s_{i,l}|^\alpha$, $|s_{i,l}|^\alpha (-s_{i,l}) = -|s_{i,l}|^{\alpha+1}$ and $z_{i,l} |s_{i,l}|^{\alpha-1} = |s_{i,l}|^{\alpha-1}$, while $s_{i,l} < 0$ is equivalent to $z_{i,l} < 0$ (negative and below the mean) which implies that $z_{i,l} |s_{i,l}|^{\alpha-1} (-s_{i,l}) = -|s_{i,l}|^\alpha$, $|s_{i,l}|^\alpha (-s_{i,l}) = |s_{i,l}|^{\alpha+1}$ and $z_{i,l} |s_{i,l}|^{\alpha-1} = -|s_{i,l}|^{\alpha-1}$. Therefore, we have $\sum_{l \neq j}^n z_{i,l} |s_{i,l}|^{\alpha-1} (-s_{i,l}) = \sum_{l \neq j}^n |s_{i,l}|^\alpha$, $\sum_{l \neq j}^n |s_{i,l}|^\alpha (-s_{i,l}) = -\sum_{l \neq j \wedge +}^n |s_{i,l}|^{\alpha+1} + \sum_{l \neq j \wedge -}^n |s_{i,l}|^{\alpha+1}$ and $\sum_{l \neq j}^n z_{i,l} |s_{i,l}|^{\alpha-1} = \sum_{l \neq j \wedge +}^n |s_{i,l}|^{\alpha-1} - \sum_{l \neq j \wedge -}^n |s_{i,l}|^{\alpha-1}$, where the indices of summation $l \neq j \wedge +$ and $l \neq j \wedge -$ denote the summation over the positive and the negative terms different from j , respectively. Since the denominator of expression (6) is strictly positive the

numerator determines the sign of $\partial \bar{s}_i^w / \partial s_{i,j}$. After having replaced these equalities, the numerator of expression (6) becomes:

$$\begin{aligned} & \left(\alpha \frac{1}{n} \sum_{l \neq j}^n |s_{i,l}|^\alpha + \sum_{l \neq j}^n |s_{i,l}|^\alpha \right) \sum_{l \neq j}^n |s_{i,l}|^\alpha \\ & + \alpha \frac{1}{n} \left(- \sum_{l \neq j \wedge +}^n |s_{i,l}|^{\alpha+1} + \sum_{l \neq j \wedge -}^n |s_{i,l}|^{\alpha+1} \right) \left(\sum_{l \neq j \wedge +}^n |s_{i,l}|^{\alpha-1} - \sum_{l \neq j \wedge -}^n |s_{i,l}|^{\alpha-1} \right), \end{aligned}$$

for all $i \in I$ and $j \in J$. The component in the first line is strictly positive. Therefore, the only way for this expression to be negative is when the term in the second line is sufficiently negative. We show that such is impossible. Consider the worst case scenario in which there is only one large positive term and many small negative terms such that $\sum_{l \neq j \wedge +}^n |s_{i,l}|^{\alpha+1} = |s_{i,l}|^{\alpha+1} > 0$ and $\sum_{l \neq j \wedge -}^n |s_{i,l}|^{\alpha+1} \approx 0$, respectively. The objective of this assumption is to obtain the largest negative term in the second line. In this case we obtain that:

$$\begin{aligned} \frac{\partial \bar{s}_i^w}{\partial s_{i,j}} & \rightarrow \frac{\left(\alpha \frac{1}{n} |s_{i,l}|^\alpha + |s_{i,l}|^\alpha \right) |s_{i,l}|^\alpha - \alpha \frac{1}{n} |s_{i,l}|^{\alpha+1} |s_{i,l}|^{\alpha-1}}{(n-1)(|s_{i,l}|^\alpha)^2} \\ & = \frac{\left(\alpha \frac{1}{n} |s_{i,l}|^{2\alpha} + |s_{i,l}|^{2\alpha} \right) - \alpha \frac{1}{n} |s_{i,l}|^{2\alpha}}{(n-1)(|s_{i,l}|^\alpha)^2} = \frac{1}{(n-1)}, \end{aligned}$$

for all $i \in I$ and $j \in J$, which is strictly positive for any profile of grades. The symmetric worst case scenario with a single negative term and many small positive terms follows the same argument. For values of $s_{i,j}$ distant from the other judges mean we show through a numerical example the failure of monotonicity. Column (6) of Table 1 shows that while the grade of the first judge moves from $s_{i,1} = 8$ to $s_{i,1} = 9$, i.e., it gets more distant from the other judges mean $\sum_{l=2}^n s_{i,l}/4 = (6+7+7+8)/4 = 7$, the WAF falls from $\bar{s}_i^w = 7.314$ to $\bar{s}_i^w = 7.303$, showing that monotonicity is not guaranteed for $s_{i,1}$ sufficiently away from the other judges mean.

Finally, following Property 4, if there exist only one grade different than $s_{i,j}$ then we can construct a situation in which occurs a downward discontinuity in the weight of judge j after an increase in her grade by an infinitesimal amount. In order to complete the proof we must verify this case. In order to construct such case assume that initially all judges (including judge j) are awarding the mean grade $s_{i,j} = \bar{s}_i = s_{i,m}$. Consequently, all grades are equally weighted by $1/n$ and $\bar{s}_i^w = \bar{s}_i = s_{i,m}$. Now, suppose that judge j increase her grade from $s_{i,m}$ to $s_{i,j} > s_{i,m}$. In this case she departs from the other $n-1$ judges and the weight $w_{i,j}$ given to her grade falls discontinuously from $1/n$ to $w^{\min} = 1/((n-1) + (n-1)^\alpha)$ (see scenario (i) in the proof of Property 4) while the weight of the other $n-1$ judges jumps from $1/n$ to

$(1 - w^{\min})/(n - 1)$. Therefore, in the discontinuous case, the difference between the new WAF \bar{s}_i^w and the initial WAF $\bar{s}_i^w = s_{i,m}$ is given by:

$$\begin{aligned}\bar{s}_i^w - s_{i,m} &= \frac{1}{(n - 1) + (n - 1)^\alpha} s_{i,j} + \left(1 - \frac{1}{(n - 1) + (n - 1)^\alpha}\right) s_{i,m} - s_{i,m} \\ &= \frac{s_{i,j} - s_{i,m}}{(n - 1) + (n - 1)^\alpha},\end{aligned}$$

which is strictly positive for all $s_{i,j} > s_{i,m}$. The proof of the case $s_{i,j} < \bar{s}_i$ follows from symmetry (Property 1). ■

Proof of Property 7. If the distribution of the grades has positive (negatively, respectively) $\alpha + 1$ -th absolute central moment, defined as $\sum_{k=1}^n |s_{i,k} - \bar{s}_i|^\alpha (s_{i,j} - \bar{s}_i)/n$ (where if $\alpha > 0$ is even and integer we have the usual $\alpha + 1$ -th central moment), then $\bar{s}_i^w < \bar{s}_i$ ($\bar{s}_i^w > \bar{s}_i$, respectively). In order to show the relation between the $\alpha + 1$ -th absolute central moment, the skewness and the WAF manipulate the inequality $\bar{s}_i^w < \bar{s}_i$ to obtain:

$$\begin{aligned}\bar{s}_i > \bar{s}_i^w &= \sum_{j=1}^n \frac{\sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha}{(n - 1) \sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha} s_{i,j} \\ &\Leftrightarrow 0 > \sum_{j=1}^n \frac{\sum_{l \neq j}^n |s_{i,l} - \bar{s}_i|^\alpha + |s_{i,j} - \bar{s}_i|^\alpha - |s_{i,j} - \bar{s}_i|^\alpha}{(n - 1) \sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha} (s_{i,j} - \bar{s}_i) \\ &\Leftrightarrow \sum_{j=1}^n |s_{i,j} - \bar{s}_i|^\alpha (s_{i,j} - \bar{s}_i) / ((n - 1) \sum_{l=1}^n |s_{i,l} - \bar{s}_i|^\alpha) > 0,\end{aligned}$$

for all $i \in I$. The sign of the expression in the numerator (i.e., the $\alpha + 1$ -th absolute central moment) determines the skewness, which is positive if $\bar{s}_i^w < \bar{s}_i$. The case $\bar{s}_i < \bar{s}_i^w$ follows the same argument. ■

References

- Arrow, K. J., 1950. A difficulty in the concept of social welfare. *The Journal of Political Economy* 58 (4), 328–346.
- Ashenfelter, O., Quandt, R., 1999. Analyzing a wine tasting statistically. *Chance* 12 (3), 16–20.
- Balinski, M., Laraki, R., 2007. A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences* 104 (21), 8720–8725.
- Balinski, M., Laraki, R., 2011. Election by majority judgment: experimental evidence. In: *In situ and laboratory experiments on electoral law reform*. Springer, pp. 13–54.

- Balinski, M., Laraki, R., 2014. Judge: Don't vote! *Operations Research* 62 (3), 483–511.
- Balinski, M. L., Laraki, R., 2010. *Majority judgment: measuring, ranking, and electing*. MIT press.
- Beliakov, G., Pradera, A., Calvo, T., 2007. *Aggregation functions: a guide for practitioners*. Vol. 221. Springer.
- Borda, J., 1784. *Histoire de l'académie royale des sciences*, 657–665.
- Brams, S. J., Fishburn, P. C., 1978. Approval voting. *American Political Science Review* 72 (03), 831–847.
- Buchanan, J. T., Henig, E. J., Henig, M. I., 1998. Objectivity and subjectivity in the decision making process. *Annals of Operations Research* 80 (0), 333–345.
- Chebotarev, P. Y., Shamis, E., 1998. Characterizations of scoring methods for preference aggregation. *Annals of Operations Research* 80, 299–332.
- Condorcet, J., 1785. *Essai sur l'application de l'analyse á la probabilité des décisions rendues á la pluralité des voix*. l'Imprimerie Royale, Paris.
- Dawes, R. M., Faust, D., Meehl, P. E., 1989. Clinical versus actuarial judgment. *Science* 243 (4899), 1668–1674.
- Emerson, J. W., Seltzer, M., Lin, D., 2009. Assessing judging bias: An example from the 2000 olympic games. *The American Statistician* 63 (2), 124–131.
- Felsenthal, D. S., Machover, M., 2008. The majority judgement voting procedure: a critical evaluation. *Homo Oeconomicus* 25 (3/4), 319–334.
- Findlay, L. C., Ste-Marie, D. M., et al., 2004. A reputation bias in figure skating judging. *Journal of Sport and Exercise Psychology* 26 (1), 154–166.
- Fishburn, P. C., Brams, S. J., 1983. Paradoxes of preferential voting. *Mathematics Magazine* 56 (4), 207–214.
- Fritz, C., Curtin, J., Poitevineau, J., Morrel-Samuels, P., Tao, F.-C., 2012. Player preferences among new and old violins. *Proceedings of the National Academy of Sciences* 109 (3), 760–763.
- Gibbard, A., 1973. Manipulation of voting schemes: a general result. *Econometrica* 41 (4), 587–601.
- Ginsburgh, V. A., Van Ours, J. C., 2003. Expert opinion and compensation: Evidence from a musical competition. *The American Economic Review* 93 (1), 289–296.
- Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E., 2011a. Aggregation functions: construction methods, conjunctive, disjunctive and mixed classes. *Information Sciences* 181 (1), 23–43.
- Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E., 2011b. Aggregation functions: means. *Information Sciences* 181 (1), 1–22.

- Hodgson, R. T., 2008. An examination of judge reliability at a major us wine competition. *Journal of Wine Economics* 3 (2), 105–113.
- Kahneman, D., Tversky, A., 1996. On the reality of cognitive illusions.
- Kingstrom, P. O., Mainstone, L. E., 1985. An investigation of the rater-ratee acquaintance and rater bias. *Academy of Management Journal* 28 (3), 641–653.
- Krantz, D. H., Luce, R. D., Suppes, P., Tversky, A., 1971. *Foundations of Measurement (Additive and Polynomial Representations)*, vol. 1. Academic Press, New York.
- Laplace, P., 1820. *Œuvres complètes de Laplace tome 7 (3rd Ed)*, v and clii–cliii.
- Lee, J., 2008. Outlier aversion in subjective evaluation. *Journal of Sports Economics* 9 (2), 141–159.
- Meehl, P. E., 1954. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.
- Miller, N., Resnick, P., Zeckhauser, R., 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51 (9), 1359–1373.
- Prelec, D., 2004. A bayesian truth serum for subjective data. *Science* 306 (5695), 462–466.
- Satterthwaite, M. A., 1975. Strategy-proofness and arrow’s conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10 (2), 187–217.
- Schlag, K. H., Tremewan, J., Van der Weele, J. J., 2015. A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics* 18 (3), 457–490.
- Smith, J. H., 1973. Aggregation of preferences with variable electorate. *Econometrica* 41 (6), 1027–1041.
- Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185 (4157), 1124–1131.
- Zitzewitz, E., 2006. Nationalism in winter sports judging and its lessons for organizational decision making. *Journal of Economics & Management Strategy* 15 (1), 67–99.