Performance Evaluation: Subjectivity, Bias and Judgment Style in Sport

António Osório[†]

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

Abstract

The number of situations that require individual judgments and evaluations, and that may be object of different sources of conscious and unconscious biases is endless. This paper proposes a practical score aggregation procedure that attempts to reduce and mitigate the influence of bias in subjective judgments. The argument is based on the idea that bias is associated with deviations from the panel mean and/or deviations from the judges' grading style. Consequently, the procedure is not specific to a particular type of bias, but rather addresses general forms of bias. We also discuss a set of desirable properties. The proposed score aggregation procedure is then applied to a unique data set from the 2000 Summer Olympic Games diving competition.

Keywords: Scores aggregation; Subjective judgments; Biased judgments; Bias correction; Sports data.
JEL classification: D71, D72, D91.

1. Introduction

One-third of all sports registered with the International Olympic Committee rely on judges' evaluations about the athletes' performance (e.g., gymnastics, diving, skating, boxing or dressage, among others). These judgments are made by qualified, but potentially biased judges. For instance, judges' decisions in sports can be biased by the existence of relationships of friendship, personal interests, social/crowd pressure, lobbies and exchange of favors, or even by the race, gender, nationality or reputation of the athlete (Section 2 discusses the different biases in sports in greater detail). In this context, it is difficult to infer whether a particular judgment is biased or not, because evaluations of qualitative performance are complex and inherently open to subjectivity and manipulation. At the same time, the technical, artistic complexity and subjectivity of sports performance, together with aspects such as time pressure

and cognitive load, make performance evaluation a difficult task (Plessner and Haar, 2006).

Performance evaluation depends on the alignment of material incentives (Baker, 1992; Dohmen and Sauermann, 2016), but also on the social environment and cognitive biases, i.e., perceptual distortion, and inaccurate or illogical interpretations (Asch, 1951; Baron, 2007; Deutsch and Gerard, 1955; Kahneman and Tversky, 1972). These biases arise from various processes that operate simultaneously in the judges' minds, such as for instance, heuristics (Shah and Oppenheimer, 2008; Tversky and Kahneman, 1974), noise and limited information processing capacity (Hilbert, 2012; Simon, 1955), emotional and moral motivations (Pfister and Böhm, 2008) and social influence (Wang et al., 2001). According to Tversky and Kahneman (1974) and Kahneman and Tversky (1996) biases can also be shortcut strategies to processing complex information.

In order to deal with these difficulties, the International Olympic Committee, together with the majority of the recognized international federations, establishes that the final score is the arithmetic mean of the scores of all the judges on the panel, and in some cases the most extreme scores are removed from the calculation (i.e., the highest and the lowest scores). However, this commonly used procedure may not be effective, since most forms of bias are subtle and can be dissimulated in a strategic way (Bassett Jr and Persky, 1994; Osório, 2017; Plessner and Haar, 2006; Wu and Yang, 2004).

In this context, we are tempted to think that a score closer to the mean is less likely to be biased. However, there is a crucial aspect to take into consideration; such a score may not be compatible with the grading style of that particular judge. For instance, if the judge in question is known to be particularly strict, i.e., to be a judge who usually awards scores well below the panel mean, a score closer to the mean may actually carry bias. In other words, the judge might be strategically hiding bias by grading closer to the panel's mean and deviating from their own style. For that reason, any score aggregation procedure must take into consideration each judge's grading style and any deviations from it. The argument can be reversed, in the sense that a score well above (or below) the mean may not necessarily suggest the existence of bias, since it may actually be compatible with that particular judge's grading style.

In our context, a judge's grading style is a measurement based on the judge's history of past scores relative to the scores awarded by the panels in which the judge has participated. If a judge has a history of consistently grading above the panel mean, then this judge might be considered as being more lenient than other judges.

On the other hand, if a judge has a history of consistently grading below the panel mean, then this judge might be considered as being stricter than other judges. In this context, Looney (2004) points out that sport governing bodies can improve the methods of performance evaluation by considering aggregation procedures that are able to capture the grading consistency of each judge.

The objective of this paper is to propose a bias correction procedure that aggregates the grades of all the judges on the panel, and that can simultaneously control for deviations in the judges' grading styles. There are several reasons to justify the introduction of this type of operational solution in sports performance evaluation (Bassett Jr and Persky, 1994; Osório, 2017; Wu and Yang, 2004), and in other dimensions of our lives (Balinski and Laraki, 2007; Balinski and Laraki, 2010; Beliakov et al., 2007; Grabisch et al., 2011a,b). First, the functioning of our society as a whole—not only sports performance evaluation—frequently relies on the ranking of objects, places, performances, projects, ideas, policies, issues, etc.[1] In this context, the development of better evaluation procedures are of first-order relevance for numerous scientific, academic and professional fields. Second, strategic bias is difficult to identify by third party monitoring. Spectators, sport governing bodies and even experts fail to detect bias and are not aware of subtle aspects like each judge's grading style. In most cases, information about each judge's grading style is not even available. Third, from a cognitive perspective, the consideration of such strategic aspects is difficult because it requires the processing of large amounts of information. Fourth, nowadays the vast majority of scoring systems are completely automated, which simplifies matters and invites the use of more complex algorithms that can help reduce and mitigate the potential effects of bias. For instance, Díaz-Pereira et al. (2014) suggest the use of human motion recognition and artificial intelligence technologies in order to reduce bias and assist judges in the decision making process (see Cust et al. (2019) for a review of this literature).

In line with the previous discussion, the score aggregation procedure proposed in this paper is designed to penalize simultaneously scores deviations from the judges' grading style and scores deviations from the judges' panel mean. These deviations

[1] In this paper, we focus mostly on sports performance, but the number of situations that require individual judgments and evaluations, which can be affected by different sources of bias, is endless. The approach in this paper may be extended to these other dimensions of our lives (e.g., the rating of any kind of items, goods and services, such as for example, wines, books, films, music, policies, scientific refereeing or any kind of talent competition, as well as, tourist locations or blog comments). Nowadays, the internet is making these evaluation procedures increasingly common.

are the ones that are most likely to be biased. The argument is that if a judge favors or penalizes a particular candidate, then this judge must be grading differently from his/her grading style and/or differently from the other judges on the same panel. Consequently, such a score should receive less weight than the scores of the other judges that are being more consistent with their grading styles and with the panel mean, and vice versa. In this context, it is the information contained in the grades of the other judges and in their grading history that determines the relevance given to each score.

Subsequently, we show that the proposed score aggregation procedure satisfies a set of desirable properties, and we consider its application to a unique data set from the 2000 Summer Olympic Games diving competitions in order to see how it reacts to the possibility of bias. We found that the implied corrections are not large enough to unequivocally support changes to the medal standings as suggested by Emerson et al. (2009). Nonetheless, the results obtained do not contradict Emerson et al. (2009). The differences are justified by the fact that the proposed score aggregation procedure corrects for the effect of deviations from the panel mean and the judges' grading style, but not so much for the influence of other forms of bias, such as for example, nationalistic bias.

To summarize, the contribution in this paper has three main aspects that distinguish it from the existing literature (see the literature review below). First, the proposed score aggregation procedure does not intend to detect and to analyze bias ex-post (i.e., after the final score is released), but to reduce and mitigate the effect of bias ex-ante (i.e., before the final score is released). Second, the proposed score aggregation procedure controls for deviations from the panel mean and/or from the judges' grading style. The consideration of deviations from the judges' grading style is new in the literature. Third, the proposed score aggregation procedure is not specific to a particular type of bias, but addresses bias in general, which makes it a useful tool for academics, practitioners and professionals in applied work. However, we must be careful, in the sense that it does not capture or remove all the existing bias and all the different types of bias. For instance, the proposed aggregation procedure has some limitations when it comes to dealing with bias that affects all or the majority of the judges, or bias towards the mean, instead of away from it, as in a Keynes beauty contest (Keynes, 1936), and it is not designed to address a particular and specific form of bias (e.g., nationalistic bias), which must be treated individually.

This paper is organized as follows: Section 2 provides a brief review of the literature. Section 3 presents the score aggregation procedure, Section 4 states and

discusses a set of desirable properties, Section 5 provides an illustrative application to the 2000 Olympic Games diving competition, and Section 6 presents the conclusions.

2. Literature review

This section reviews (i) the literature on sports performance evaluation bias with a brief reference to other cases of performance evaluation, and (ii) the literature on preferences and judgments aggregation.

In order for judges to act in accordance with the interests of the associated competition organizing body, the material incentives should be aligned (Baker, 1992; Dohmen and Sauermann, 2016). As in a principal-agent relationship, unbiased judgments should be rewarded and biased judgments should be punished. In this context, bribes, friendships, personal interests and lobbies distort incentives and consequently induce biased decisions (Duggan and Levitt, 2002; Wolfers, 2006).

However, individual decisions also depend on non-material aspects associated with the social environment and on cognitive biases, e.g., perceptual distortions, and inaccurate or illogical interpretations (Asch, 1951; Baron, 2007; Deutsch and Gerard, 1955; Kahneman and Tversky, 1972). These biases arise from various processes that operate in the judges' minds and that are difficult to separate from each other—for instance, heuristics (Shah and Oppenheimer, 2008; Tversky and Kahneman, 1974), noise and limited information processing capacity (Hilbert, 2012; Simon, 1955), emotional and moral motivations (Pfister and Böhm, 2008), and social influence (Wang et al., 2001).

The list of cognitive biases reported over the last decades is continuously evolving (Baron, 2007). In this context, the complexity of sports performance evaluation together with aspects like time pressures, cognitive load, and performance subjectivity makes this subject very active in terms of research. The following review of the literature offers a brief summary of some of this research. For more exhaustive reviews of the literature, see Bar-Eli et al. (2011), Dohmen and Sauermann (2016) and Plessner and Haar (2006).

Nationalistic bias is a particular type of bias that has been frequently reported in sports performance evaluation literature. For instance, Coupe et al. (2018) studied bias in the FIFA Ballon d'Or award for the best soccer player. They found that judges are biased towards candidates from their own country, national team, continent and league team. Popović (2000) examined the rhythmic gymnastics competition in the 2000 Summer Olympics and found that judges tend to favor their

own country's gymnasts, but not sufficiently to be statistically significant. Similarly, Zitzewitz (2006) examined the figure skating and the ski jumping competitions in the 2002 Winter Olympics and found evidence in favor of nationalistic bias (see also Lock and Lock, 2003; Zitzewitz, 2014). Emerson et al. (2009) examined the diving competition in the 2000 Summer Olympics and concluded that nationalistic bias could have influenced the final medals standing.

Similarly, using data from the Eastern Ontario and Québec sections of Skate Canada, Findlay and Ste-Marie (2004) found reputation bias in figure skating. The ranks were better when the skaters were evaluated by judges who knew their reputation than when evaluated by judges who did not knew their reputation.

In gymnastics, within-team order bias is particularly common. In this case, biased expectations are induced by the common strategy used by coaches of placing their strongest gymnasts later in the order of rotation. Plessner and Haar (2006) found that this strategy induces judges to give higher marks to performances at the end of the rotation order than if that same performance had been observed earlier in the rotation order. In the same way, Damisch et al. (2006) found that sequential performance judgments in sports are biased by the previously judged performances, which depends on the degree of perceived similarity between the successive performances.

Using data from World Figure Skating Championships between 2001 and 2003, Lee (2008) show the existence of outlier aversion bias, in which judges avoid grading far from the panel mean, as in a beauty contest (Keynes, 1936).

The home team advantage is another well studied form of bias, being observed in many sports like football, basketball, baseball and ice hockey, and is often explained by the crowd's influence on judges' decisions (Dohmen and Sauermann, 2016; Garicano et al., 2005; Nevill et al., 1996; Price et al., 2012; Sutter and Kocher, 2004; Unkelbach and Memmert, 2010). In the same vein, Page and Page (2007) found that teams have a higher chance of qualifying for the next round when they play the second leg at home.

Racial bias in sports—which is frequent in other dimensions of our lives—has been found among National Basketball Association referees (Price and Wolfers, 2010; Larsen et al., 2008), and among Major League Baseball umpires (Parsons et al., 2011).

Other forms of bias, which are not so common, have been reported in the sports literature (Dohmen and Sauermann, 2016). For instance, Helsen et al. (2006) found the existence of cognitive and perceptual biases with offside calls. Offside calls depend crucially on the position of the referee relative to the players. Frank and Gilovich (1988) found that shirt color can induce cognitive biases amongst football and ice hockey players.

In this paper, we focus on sports performance evaluation, but the number of situations that require performance evaluations, and that are affected by different sources of bias is endless. Bias is not merely an issue prevalent in subjective evaluations, but inherent to every dimension of life (Buchanan et al., 1998).[2] The score aggregation procedure in this paper attempts to mitigate the effect of bias from performance evaluation.

In addition to the limitations associated with subjective judgments, there are also difficulties at the aggregation stage. A large body of literature in sports performance evaluation (Bassett Jr and Persky, 1994; Osório, 2017; Wu and Yang, 2004), and judgment in general (Balinski and Laraki, 2007; Balinski and Laraki, 2010; Balinski and Laraki, 2014; Felsenthal and Machover, 2008), has proposed different solutions to aggregate the preferences of different individuals (Beliakov et al., 2007; Grabisch et al., 2011a,b). For instance, Osório (2017) proposes an aggregation procedure that corrects deviation from the panel mean, while this paper goes a step further, by proposing an aggregation procedure that can also correct deviations from each judge's grading style.

The most common solution, among the International Olympic Committee and the international federations, is "range voting", in which judges rate the candidates with a grade within a specified interval. The candidate with the highest sum or average wins. The method is easy to implement and passes certain generalizations of the Arrow (1950) impossibility theorem, but it is particularly sensitive to bias and strategic manipulation.

Often, in order to deal with this difficulty a truncation is used to remove extreme scores and mitigate potential bias. In this context, "majority judgment" ranks candidates by the median score, i.e., all scores are truncated, except the middle one, which becomes the final score (Balinski and Laraki, 2007; Balinski and Laraki, 2010). This procedure is more robust to manipulation and reduces the incentives to exaggerate.

[2]Several studies have focused on bias in evaluation contexts other than sports. For instance, to mention just a few, in musical competitions, Ginsburgh and Van Ours (2003) found that judging panel members are influenced by the order of appearance of candidates, while Tsay (2013) found that judges are influenced more by what they see than by what they hear. In the Eurovision Song Contest, Ginsburgh and Noury (2008) found that linguistic and cultural similarities between singers and judges are determinant, while in academic awards, Hamermesh and Schmidt (2003) found that affiliation is crucial in the judges' decision. In this context, some statistically based rating procedures have shown better results than expert opinions (Dawes et al., 1989; Meehl, 1954). Other inconsistencies and paradoxical observations are reported in the literature (Ashenfelter and Quandt, 1999; Fritz et al., 2012; Hodgson, 2008; Plessner and Haar, 2006). Further development of these issues is beyond the scope of the present paper.

However, excessive truncation leads to a loss of information and diversity, in particular if bias is only a possibility. The score aggregation procedure proposed in this paper preserves the information and diversity of opinions contained in the judges panel while mitigating the effects of bias.

3. The score aggregation procedure

In general, there is no evidence to prove conclusively whether a particular score is biased or not. Moreover, it is virtually impossible to control all forms of conscious and unconscious bias and manipulation. Another difficulty is that the judges' preferences and interests are private information and impossible to determine ex-ante. In this context, any score aggregation procedure must depend only on what is known, which in many cases is not too much. In what follows, we propose a score aggregation procedure that attempts to deal with these practical limitations and to mitigate the effect of bias.

In this context, we control for bias in two dimensions. The first dimension controls for score deviations from the judges' historical grading style. Each judge has a unique grading style. Some judges are systematically more strict or lenient than others. The second dimension controls for score deviations from the panel of judges' mean score.

Let $s_{ij} \in [S_-, S_+] \subset \mathbb{R}$ be the score awarded by judge $j \in J = \{1, ..., n\}$ for the performance of competitor $i \in I = \{1, ..., m\}$. We consider well-defined scores on numerical scales, with no language-consistency issues among the judges, e.g., $[S_-, S_+] = [0, 10]$. Let $\mathbf{s}_i = (s_{i1}, ..., s_{in})$ denote the vector of scores awarded by the panel of judges for the performance of competitor $i \in I$. The mean score of the performance of the competitor $i \in I$ is denoted as $\overline{s}_i$ and corresponds to the arithmetic mean over the scores awarded by all judges, i.e., $\overline{s}_i \equiv \frac{1}{n} \sum_{j=1}^{n} s_{ij}$.

In addition, in order to determine each judge's grading style, we consider the history of past scores. Let the history of the past scores awarded by the judge $j \in J$ be denoted as $\mathbf{h}_j^t$, and let the history of past mean scores awarded on the panels on which judge $j \in J$ participates be denoted as $\mathbf{h}_{(j)}^t$, where the superscript "t" denotes the moment in time. The vectors $\mathbf{h}_j^t$ and $\mathbf{h}_{(j)}^t$ consist of the past scores that are considered relevant in defining the judge's j style, i.e., $\mathbf{h}_j^t = (s_{.j}^{t-1}, s_{.j}^{t-2}, ...)$ and $\mathbf{h}_{(j)}^t = (\overline{s}_{.(j)}^{t-1}, \overline{s}_{.(j)}^{t-2}, ...)$, respectively, where the subscript "." expresses the irrelevance of the competitors identity associated with that history of past scores. For example, these vectors may consist of all the scores awarded over the last year, or all the scores

awarded up to the present event, or any other criteria.[3] In order to keep the notation as simple as possible, when possible, we remove the explicit reference to time "t".

In this context, in order to determine each judge's grading style, one possibility is to aggregate the history of past scores into a single measure (e.g., a simple average, a weighted average, or any other stable criteria).[4] Let $\overline{s}_{\mathbf{h}_j} \equiv \frac{1}{T}\sum_{t=1}^{T} s_{.j}^{t}$ be the arithmetic mean of judge $j's \in J$ history of past scores (where T is the number of scores considered for the history of judge $j \in J$), and $\overline{s}_{\mathbf{h}_{(j)}} \equiv \frac{1}{T}\sum_{t=1}^{T} \overline{s}_{.(j)}^{t}$ be the arithmetic mean of the history of past mean scores awarded by the panels on which judge $j \in J$ has been involved (i.e., the mean of the history of panel means).[5] In our context, given the performance of competitor i and the panel mean $\overline{s}_i$, the style adjusted expected grade of judge $j \in J$ to competitor i, i.e., the grade of judge j that would be compatible with his/her own style, is defined as follows.

Definition 1. *The style adjusted expected grade of judge j is defined as*

$$\overline{s}_{i\mathbf{h}_j} = \overline{s}_i \overline{s}_{\mathbf{h}_j} / \overline{s}_{\mathbf{h}_{(j)}},$$

where the ratio $\overline{s}_{\mathbf{h}_j}/\overline{s}_{\mathbf{h}_{(j)}}$ defines the grading style of judge j.

---

[3]The grading style must not vary with the order in which the history is presented. This aspect has implications for Section 4, when we discuss some of the properties of the proposed aggregation procedure. This implies that the history cannot include scores entered in the present competition. Otherwise, the order of the performances could interfere with the measurement of the grading style, which must be stable throughout the competition. This aspect also places restrictions on the use of moving averages as aggregate measurements of the history.

[4]We are intentionally ambiguous about the length of the history of past scores. We leave this decision to the social planner or the sport's governing body responsible for the competition. The longer the history and the closer in time the better. However, on the same panel, we may have judges with different histories in terms of length, but also in terms of quality. In that sense, homogenizing all the histories by using the length of the shortest history as a reference may not be a good idea, because it could imply a loss of data about the history of the other judges.

[5]Grading style can be defined in different ways. These alternatives have in common the use of information from the history of past scores. For instance, grading style could have been defined as:

$$\overline{s}_{i\mathbf{h}_j} = \overline{s}_{\mathbf{h}_j} \overline{s}_i / \overline{s}_{\mathbf{h}},$$

where $\overline{s}_{\mathbf{h}} \equiv \frac{1}{n}\sum_{j=1}^{n} \overline{s}_{\mathbf{h}_j}$ is the arithmetic mean of the grading histories of all the judges on the panel. The results in this paper would not change significantly if we were to consider this measure. However, grading style is more correctly defined if judge $j's$ grading history is made relative to the mean score of the panels on which judge j has participated. Alternatively, grading style could have been defined as: $\overline{s}_{i\mathbf{h}_j} = \overline{s}_i \sum_{t=1}^{T}(s_{.j}^{t}/\overline{s}_{.(j)}^{t})$. This case is conceptually equivalent to the one in this paper, and leads to almost exactly the same results, but the approach in this paper is more intuitive and simpler to apply. Other definitions of grading style are also possible.

Therefore, the ratio $\overline{s}_{\mathbf{h}_j}/\overline{s}_{\mathbf{h}_{(j)}}$ captures judge $j's$ history of deviations from the panel mean, and defines judge $j's$ grading style.[6]

The following example provides an illustration.

Example 1. *Suppose judges* 1, 2 *and* 3 *awarded the scores* $s_{i1} = 7.00$, $s_{i2} = 7.00$ *and* $s_{i3} = 8.00$, *respectively, with* $\overline{s}_i = 7.33$. *Suppose also that their history is summarized by the mean of the past scores, i.e.,* $\overline{s}_{\mathbf{h}_1} = 8.00$, $\overline{s}_{\mathbf{h}_2} = 7.00$ *and* $\overline{s}_{\mathbf{h}_3} = 6.00$, *respectively, and by the mean of the past means awarded by the panels on which these judges have been involved, i.e.,* $\overline{s}_{\mathbf{h}_{(1)}} = 7.00$, $\overline{s}_{\mathbf{h}_{(2)}} = 7.00$ *and* $\overline{s}_{\mathbf{h}_{(3)}} = 7.00$, *respectively. Then, each judge style adjusted expected grade would be given by* $\overline{s}_{i\mathbf{h}_1} = 8.38$, $\overline{s}_{i\mathbf{h}_2} = 7.33$ *and* $\overline{s}_{i\mathbf{h}_3} = 6.29$, *respectively.*

This example suggests that despite judges 1 and 2 seeming to agree on a final score of 7.00, judge 3, by proposing a score of 8.00, might be deviating from his/her grading style. Note that judge 3 has a history of being strict by awarding on average $\overline{s}_{\mathbf{h}_3} = 6.00$ on panels that awarded on average $\overline{s}_{\mathbf{h}_{(3)}} = 7.00$. In this context, in order to be consistent with his/her grading style and the scores of the other judges, judge 3 should have proposed a score somewhere near 6.86. The scores aggregation procedure proposed in this paper has the objective of reducing the influence of diverging scores like the score awarded by judge 3. However, we must be careful, because bias is only a possibility.

In the context of the present paper, this objective will be achieved by reducing the weight given to the divergent scores. For that reason, the scores aggregation procedure proposed in this paper will give a weight of 33.8% and 47.5% to the scores of each of the judges 1 and 2, respectively, and only a weight of 18.7% to the score of judge 3 (for $\alpha = 1/2$ and $\gamma = 2$, see below).

Formally, the weights are functions $w_{ij} : D_1 \times ... \times D_n \to [0, 1]$, where $D_k = [S_-, S_+]^{1+|\mathbf{h}_k|+|\mathbf{h}_{(k)}|}$, $|\mathbf{h}_k|$ denotes the cardinality of each judge $k's$ history of past scores, and $|\mathbf{h}_{(k)}|$ denotes the cardinality of the history of past mean scores awarded on the panels participated in by judge k (note that we may have $|\mathbf{h}_k| = |\mathbf{h}_{(k)}|$). In other words, the weights depend on the history of past scores of each judge $j \in J$, i.e.,

[6] The history of past grades of each judge usually includes scores from different competitions with different levels and different stages. In this context, for instance, the average score in the early stages of the same competition tends to be lower than the average scores in the later stages, in which only the best competitors are left . Similarly, the average scores in national competitions tend to be lower than the average scores in the Olympics, because in the Olympic Games competitors tend to be better on average. The ratio $\overline{s}_{\mathbf{h}_j}/\overline{s}_{\mathbf{h}_{(j)}}$ corrects for this heterogeneity.

$\{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n$, which defines judge $j's$ grading style $\overline{s}_{i\mathbf{h}_j}$, and on the scores awarded by all the judges, i.e., the vector $\mathbf{s}_i$.

Definition 2. *The weights are defined as:*

$$w_{ij}(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) \equiv \frac{\sum_{k\neq j}^n (\alpha\,|s_{ik} - \overline{s}_{i\mathbf{h}_k}| + (1-\alpha)\,|s_{ik} - \overline{s}_i|)^\gamma}{(n-1)\sum_{k=1}^n (\alpha\,|s_{ik} - \overline{s}_{i\mathbf{h}_k}| + (1-\alpha)\,|s_{ik} - \overline{s}_i|)^\gamma}, \tag{1}$$

for all $j \in J$.

Consequently, given the performance of competitor $i \in I$, the vector of scores awarded by the n judges are aggregated into a single score, according to the following definition.

Definition 3. *The score aggregation procedure is defined as:*

$$\overline{s}_i^*(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) \equiv \sum_{j=1}^n w_{ij} s_{ij}, \tag{2}$$

where w_{ij} *represents the weight given to the score* s_{ij} *awarded by judge* $j \in J$ *for the performance of competitor* $i \in I$, *with* $\sum_{j=1}^n w_{ij} = 1$.

In case of a tie between two or more competitors, the reader is free to consider any tie-breaking rule.

Definitions 1, 2 and 3 fully describe the score aggregation procedure proposed in this paper.

The parameters in Definition 2 have the following interpretation. The parameter $\alpha \in [0, 1]$ controls the importance given to score deviations from the judges grading style, and $1 - \alpha$ controls the importance given to score deviations from the panel mean.[7] In the particular case where $\alpha = 1$, only the deviations from the grading style are punished, while in the particular case where $\alpha = 0$, only the deviations from

[7] We can consider alternative weight functions, but with similar implications. For instance, we can consider different parameters to control for deviations from judge $j's$ grading style and from the panel mean, i.e., β and γ, respectively. In this case, we could have:

$$w_{ij}(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) \equiv \frac{\alpha \sum_{k\neq j}^n |s_{ik} - \overline{s}_{i\mathbf{h}_k}|^\beta + (1-\alpha)\sum_{k\neq j}^n |s_{ik} - \overline{s}_i|^\gamma}{(n-1)(\alpha \sum_{k=1}^n |s_{ik} - \overline{s}_{i\mathbf{h}_k}|^\beta + (1-\alpha)\sum_{k=1}^n |s_{ik} - \overline{s}_i|^\gamma)}.$$

Other formulations are also possible. For instance, we can also consider the following simplified

the panel mean are punished. However, since we are interested in punishing both types of deviations simultaneously, we set $\alpha \in (0,1)$.

The parameter $\gamma \geqslant 0$ determines the magnitude of punishment of the score deviations. The larger the value of γ, the stronger the punishment to scores that are distant from the judges' grading style and from the panel mean. However, values of γ that are too large can be problematic because bias is only a possibility, and we do not want to distort the results in cases in which there is no bias. On the other hand, low values of γ may not penalize biased scores enough.[8]

In our context, Expression (1) is written in the most general form, and the parameters α and γ are controlled by the social planner or the sport's governing body. However, in applied and operational work, in order to simplify the analysis, we can consider $\alpha = 1/2$ (i.e., equal importance to both types of deviations) and $\gamma = 2$ (i.e., the quadratic distance). The proposed score aggregation procedure is particularly flexible when it comes to accommodate different possibilities.

The weight given to judge j in Expression (1) increases with the deviations of the other judges k from their grading style $|s_{ik} - \overline{s}_{i\mathbf{h}_k}|$ and the panel mean $|s_{ik} - \overline{s}_i|$, and decreases with the deviations from the judge $j's$ own grading style $\left|s_{ij} - \overline{s}_{i\mathbf{h}_j}\right|$ and the panel mean $|s_{ij} - \overline{s}_i|$. Expression (1) also meets the objective of penalizing most the largest score deviations from the judges' grading style and from the panel mean, which are the scores that are most likely to be biased. Simultaneously, the correction mechanism gives more weight to judges that are assumed not to be biased, which are the ones whose scores show higher prevalence and similitude, and are more consistent with their own grading style. This idea motivates the score aggregation procedure in this paper.

Intuitively, on the one hand, the first term in the numerator of Expression (1)

---

weighted mean formulation:

$$w_{ij}(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) \equiv \frac{\sum_{k\neq j}^n |s_{ik} - (\alpha\overline{s}_{i\mathbf{h}_k} + (1-\alpha)\overline{s}_i)|^\gamma}{(n-1)\sum_{k=1}^n |s_{ik} - (\alpha\overline{s}_{i\mathbf{h}_k} + (1-\alpha)\overline{s}_i)|^\gamma}.$$

These formulations may differ slightly in terms of properties, but the crucial aspect is that all of them penalize deviations from the judges grading style. Other approaches, like majority judgment, which is based on the median score, have also been considered in the literature (Balinski and Laraki, 2007; Balinski and Laraki, 2010; Bassett Jr and Persky, 1994; Wu and Yang, 2004)

[8] In the case where $\gamma \to 0$, the score aggregation procedure converges to the mean $\overline{s}_i^* \to \overline{s}_i$, because all grades are equally weighted, while in the case where $\gamma \to \infty$, the score aggregation procedure ignores the most extreme score and weights all the other scores equally (with some specificities in the case of more than one extreme score).

considers score deviations from the grading style of all the judges other than judge $j \in J$, i.e., all $k \neq j \in J$. On the other hand, the second term in the numerator of Expression (1) considers score deviations from the panel mean of all the judges other than judge $j \in J$, i.e., all $k \neq j \in J$. Therefore, the more (respectively, the less) judge $j \in J$ deviates from his/her grading style and from the panel mean, the more (respectively, the less) weight receives the scores of the other judges $k \neq j \in J$, and the less (respectively, the more) weight receives his/her scores.

In this context, in order to understand the intuition behind the proposed aggregation method, consider the following decomposition of judge $j's$ evaluation of competitor $i's$ performance:

$$s_{ij} = a_i + u_{ij} + b_{ij},$$

where a_i is the unknown actual evaluation of competitor $i's$ performance, u_{ij} is the unbiased deviation of judge j, which is i.i.d for each judge according to some distribution, and b_{ij} is the subjective bias of judge j towards competitor i, which can also be seen as a random variable. Therefore, u_{ij} captures judge $j's$ grading style, while b_{ij} captures judge $j's$ subjective bias, where we are assuming that u_{ij} and b_{ij} are independent and that the proposed additive decomposition exists. In this context, judge $j's$ grading style component of s_{ij}, in the absence of bias, is equal to $\overline{s}_{i\mathbf{h}_j} = a_i + u_{ij}$, while the panel mean is equal to $\overline{s}_i = \sum_{k\neq j}^{n}(a_i + u_{ik} + b_{ik})/n = a_i + \sum_{k\neq j}^{n} u_{ik}/n + \sum_{k\neq j}^{n} b_{ik}/n$. Therefore, the distinctive component of judge $j's$ weight in Expression (1) can be written as:

$$\alpha \left|s_{ij} - \overline{s}_{i\mathbf{h}_j}\right| + (1-\alpha)\left|s_{ij} - \overline{s}_i\right| = \alpha \left|b_{ij}\right| + (1-\alpha)\left|u_{ij} - \sum\nolimits_{k\neq j}^{n} u_{ik}/n + b_{ij} - \sum\nolimits_{k\neq j}^{n} b_{ik}/n\right|.$$

In other words, the deviations from the grading style capture the subjective bias of judge j, while the deviations from the panel mean capture deviations from the other judges' average grading style and average subjective bias, respectively.

This decomposition provides an alternative intuition into how the proposed aggregation method mitigates and reduces the effects of subjective bias, i.e., either directly, by means of subjective bias itself, or indirectly, by means of deviations from the other judges' subjective bias.

Figure 1, which is in connection with Example 1, illustrates the essence of the score aggregation procedure, for the case of three judges and when the score of judge 3 varies. Briefly, on the left-hand side of Figure 1, since judge 2 is grading nearer his/her style and the panel mean than the other judges, the weight given to judge 2 always remains high. Simultaneously, the weight given to judge 3 decreases as

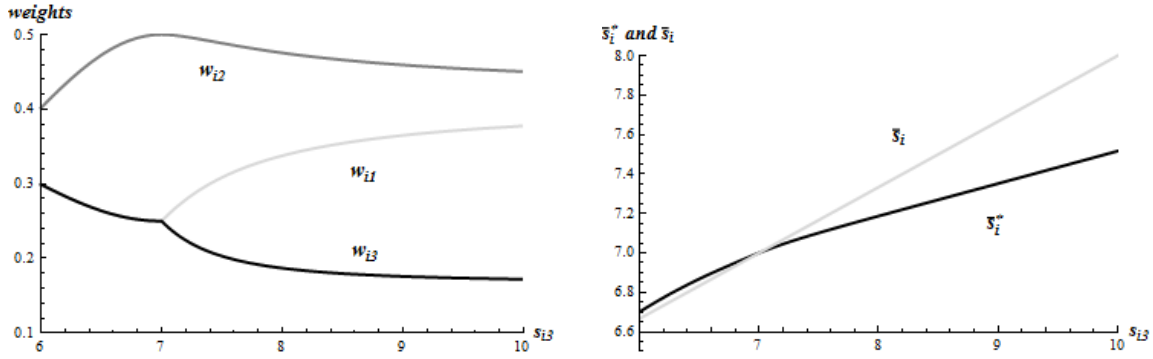

Figure 1: On the left-hand side - judges 1, 2 and 3 weights w_{ij} for varying s_{i3}. On the right-hand side - the score aggregation function $\overline{s}_i^*$ and the arithmetic mean $\overline{s}_i$ for varying s_{i3}. (vector of scores $(7.00, 7.00, s_{i3})$, vector of mean past scores $(8.00, 7.00, 6.00)$, and vector of mean past panel mean scores $(7.00, 7.00, 7.00)$, for $\alpha = 1/2$ and $\gamma = 2$)

he/she moves away from his/her grading style and the panel mean. On the other hand, the weight given to judge 1 increases as judge 3 grades above 7.00, because in relative terms, the score of judge 1 is becoming more consistent with his/her grading style and the panel mean. Consequently, the right-hand side of Figure 1 shows the decreasing impact of judge $3's$ score on the final score as the mechanism corrects the increasing deviations from his/her grading style and the panel mean.

4. Properties of the score aggregation procedure

In this section, we take a closer look at some additional properties of the proposed scores aggregation procedure in Expression (2) and its weights given by Expression (1). We adapt into our context some basic properties that have been considered in the literature (Balinski and Laraki, 2007; Felsenthal and Machover, 2008), and which are not always easily satisfied by other aggregation procedures (Beliakov et al., 2007; Grabisch et al., 2011a,b). The analysis of these properties provides a deeper understanding of the proposed score aggregation procedure that can be useful to researchers, sport-governing bodies, decision-makers and practitioners.

The proof of these properties is simple and for that reason omitted. They follow directly from the definition of the score aggregation procedure in Expression (2) and the definition of weights in Expression (1).

A commonly desired property is homogeneity of degree zero in weights, which implies that the weights are independent of the units of measurement.

Property 1 (homogeneity). *The weights are homogeneous of degree zero, i.e.,* $w_{ij}(\lambda \mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) = w_{ij}(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n)$ *for all* $i \in I$ *and* $j \in J$*, and for any* $\lambda > 0$.

This property means that if we double the score of all the judges, the weight given to each judge remains unchanged. This property is passed on to the score aggregation procedure, which becomes scale-consistent, i.e., homogeneity of degree one in the final score. Consequently, if we double the score of all the judges, the final score doubles, but the ranking of each competitor remains unchanged.

Property 2 (scale-consistent). *The score aggregation procedure is scale-consistent, i.e.,* $\overline{s}_i^*(\lambda \mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) = \lambda \overline{s}_i^*(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n)$ *for all* $i \in I$*, and for any* $\lambda > 0$.

However, in our context, the score aggregation procedure depends on the identity of each judge, because each judge has a different grading style, which is characterized by his/her history of past scores (see the discussion in Footnote 3). This aspect is crucial in order to reduce and mitigate the possible effects of bias.

The absolute value function employed in the proposed score aggregation procedure guarantees an equal treatment of scores on both sides of the judges' grading style and the panel mean. This aspect is important because bias may be hidden above or below the judges' grading style and the panel mean. Monitoring is achieved by considering simultaneously deviations from these reference values. In this context, the proposed score aggregation procedure returns the arithmetic mean when the grades of all the judges on the panel coincide.

Property 3 (unanimity). *If* $s_{ij} = s_{ik}$ *for all* $j, k \in J$*, then* $\overline{s}_i^* = \overline{s}_i$.

The property does not imply that weights are the same, because each judge has a different grading history or style, but it implies that if all judges awarded the same grade to a given performance, then the final score must be that grade. Nonetheless, we must note that a strategic judge (strict or lenient) can be hiding bias even when grading in line with all the other judges. In this particular case, the aggregation procedure reflects the difficulty in building a strong argument in the event of biased behavior.

In addition, the score aggregation procedure must be independent of irrelevant alternatives. In other words, the grades awarded to competitors other than competitor $i \in I$ cannot affect the final score of competitor $i \in I$, and the judges' past scores not considered in the history cannot affect the final score of competitor $i \in I$ (see Footnote 3).

Property 4 (independence of irrelevant alternatives). *The score aggregation procedure is independent of irrelevant alternatives, i.e.,* $\overline{s}_i^*$ *is independent of everything not in* $\mathbf{s}_i$ *and* $\{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n$.

The score aggregation procedure should also be continuous, where continuity has the usual mathematical meaning. In other words, small changes in the numerical scores (i.e., the input), should imply small changes in the final score (i.e., the output). This property is convenient for most practical applications.

Property 5 (continuity). *The score aggregation procedure* $\overline{s}_i^*(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n)$ *is continuous in* $\mathbf{s}_i$.

Note also that the score aggregation procedure is differentiable, except when the absolute value function is not differentiable, i.e., when $s_{ik} = \overline{s}_{i\mathbf{h}_k}$ or $s_{ik} = \overline{s}_i$. Differentiability almost everywhere is also a convenient property for practical applications.

Note also that in general the scores aggregation procedure $\overline{s}_i^*$ tends to be a monotonic function of s_{ij}. The exception occurs for sufficiently large score deviations from the grading style or the panel mean, and when these deviations are heavily punished (i.e., by means of a large value of γ). Therefore, the failure of this property occurs only under extreme circumstances and is due to the bias correction objective implicit in the score aggregation procedure. For instance, if a judge awards a score relatively higher than his/her grading style or the panel mean, the final score may fall if the decrease in the weight given to that judge is stronger compared to the increase in the score.

Lastly, Properties 1-5 cannot uniquely characterize the proposed scores aggregation procedure. The difficulty arises from the fact that the weights in Expression (1) are not constant and depend in a nonlinear way on the scores that they weigh.[9]

[9]The score aggregation function can be written in more general terms as:

$$\overline{s}_i^+(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) \equiv \sum_{j=1}^n f_j(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) g(s_{ij}),$$

5. A data application to the Olympic Games

In this section, we apply the proposed score aggregation procedure to the diving competition of the 2000 Summer Olympic Games. The objective is to illustrate the application of the score aggregation procedure to real data, and to discuss some implementation issues and the obtained results.

The data set is obtained from Emerson et al. (2009), and is composed of 10,788 dives with specific information about the score and the difficulty of each dive, the identity of each diver and the identity of each judge for the preliminary round, the semi-final and the final stages of the event. The level of detail in the available information makes this data set particularly unique for studying bias in sports performance evaluation.

We start by describing the aggregation procedure used by the International Olympic Committee to compute the final score. The judges awarded scores ranging from 0 to 10 in increments of 0.5. The judging panel was composed of seven judges making independent assessments about each dive quality. For each dive, the final score is calculated by removing the lowest and the highest scores and averaging the middle five scores. The scores were then multiplied by the degree of difficulty DD_i and by 3, in accordance with the following formula:

$$Olympic\ score\ (dive\ i) = DD_i \times 3 \times \overline{s}'_i, \tag{3}$$

for all $i \in I$, where $\overline{s}'_i$ denotes the truncated average resulting from the middle five scores, i.e., $\overline{s}'_i = (\sum_{j=1}^{7} s_{ij} - \min_j \{s_{ij}\} - \max_j \{s_{ij}\})/5$.

In order to compare our results with the International Olympic Committee, we also remove the lowest and highest scores. Note that in our context, in the case of more than one lowest and highest score, the removed score is the one associated with the judge with the largest deviation from his/her grading style. Therefore, we may be already removing some potentially biased scores. The final score is then obtained by multiplying the scores aggregation procedure by the degree of difficulty DD_i and by 3, as in the International Olympic Committee procedure, according to

where $f_j(.)$ is a weight function that receives the vectors $\mathbf{s}_i$ and $\{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n$ as inputs, and $g(.)$ is a function that receives the grade of judge j as input. Then, if the function $f_j(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n)$ is continuous, homogeneous of degree zero on $\mathbf{s}_i$, with $f_j(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) \in [0,1]$ and $\sum_{j=1}^n f_j(\mathbf{s}_i, \{\mathbf{h}_k, \mathbf{h}_{(k)}\}_{k=1}^n) = 1$, and the function $g(s_{ij}) = s_{ij}$, then the properties of the general aggregation function $\overline{s}_i^+$ match the properties of $\overline{s}_i^*$.

the following formula:

$$Scores\ aggregation\ procedure\ (dive\ i) = DD_i \times 3 \times \overline{s}_i^*, \tag{4}$$

for all $i \in I$, where $\overline{s}_i^*$ denotes the score aggregation procedure in Expression (2) with the weights given by Expression (1).

Lastly, in both procedures, the scores obtained in each dive are added up to obtain each diver's final score.

In what follows, we analyze the men's 3-meter springboard and the women's 10-meter platform diving competitions. In these two events, the difference between the first two divers is very narrow so that the medals final standing could have been easily influenced by the presence of bias.

5.1. ***The men's 3-meter springboard diving competition***

In the 2000 Summer Olympics, the diver Xiong Ni of China won the gold medal with an extremely narrow margin from Fernando Platas of Mexico (Column (1) of Table 1). The result generated controversy because of the eleven dives counting for the final score (i.e., six dives from the final stage and five dives from the semi-final stage), three dives were awarded by a committee with a judge from the same nationality as the winning diver. The Chinese judge Facheng Wang participated in the semi-final stage, and three of his judgments counted towards the final score. Note that judges with the same nationality as the competitors are not normally assigned to the final stage, although they can be assigned to earlier stages of the competition, as in this case to the semi-final stage.

Some years later, Emerson et al. (2009) studied the diving competition of the 2000 Summer Olympic Games. However, their results were not sufficiently significant to support the argument that the judge Facheng Wang benefited the diver Xiong Ni in the men's 3-meter springboard diving competition.

In what follows, we apply the scores aggregation procedure proposed in this paper to the 2000 Summer Olympics men's 3-meter springboard diving competition data and discuss the results obtained.

The final Olympic score calculated using (3) is shown in Column (1) of Table 1. The application of the proposed score aggregation procedure, with $\alpha = 1/3$ and $\gamma = 2$, to the grades awarded in the eleven dives returns the first place to Ni Xiong with 709.74 points against Fernando Platas with 709.33 points (Column (2) in Table 1). Similarly, the application of the proposed score aggregation procedure, with $\alpha = 2/3$ and $\gamma = 2$, to the grades awarded in the eleven dives returns the first place

DIVER	Olympics (1)	SAP $\alpha = 1/3$ (2)	SAP $\alpha = 2/3$ (3)
Xiong Ni (CHN)	708.72	709.74	710.07
Fernando Platas (MEX)	708.42	709.33	709.03
Dmitri Sautin (RUS)	703.02	704.35	704.36
Xiao Hailiang (CHN)	671.04	670.66	670.30
Dean Pullar (AUS)	647.40	647.82	647.68
Troy Dumais (USA)	642.72	641.47	641.38
Mark Ruiz (USA)	638.33	636.69	636.56
Ken Terauchi (JPN)	634.47	633.14	633.07
Stefan Ahrens (GER)	619.17	617.34	617.79
Andreas Wels (GER)	616.53	614.27	614.69
Imre Lengyel (HUN)	613.47	613.68	613.69
Tony Ally (GBR)	583.80	584.99	584.98

Table 1: **The men's 3-meter springboard diving competition final scores:** comparison of the Olympic Committee procedure (Olympics) and the scores aggregation procedure (SAP) for $\gamma = 2$. Source: Author's own elaboration with data from Emerson et al. (2009).

to Ni Xiong with 710.07 points against Fernando Platas with 709.03 points (Column (3) in Table 1).[10] Therefore, the proposed score aggregation procedure corroborates Emerson et al. (2009) and the medal's final standing.[11] However, the medal's final standing would have changed for $\alpha < 0.12$ with $\gamma = 2$.

In what follows, we discuss the results obtained and their intuition in more detail. In this context, consider the information in Tables 2 and 3 regarding the scores awarded to the divers Xiong Ni and Fernando Platas, respectively, by a panel of judges in which the judge Facheng Wang participated. The Column with the label

[10]We consider the intermediate cases $\alpha = 1/3$ and $\alpha = 2/3$ with $\gamma = 2$, because they are sufficiently representative and informative.

[11]The statistical method employed by Emerson et al. (2009) is particularly powerful for detecting bias and manipulation. However, since bias can be hidden in very complex and strategic ways, there is no perfect method to deal with this possibility. For instance, a judge may penalize a particular athlete in the early stages of competition, in which the qualification of that athlete is almost guaranteed (because of the athlete's quality), in order to later benefit this same athlete in the most crucial stages of the competition. Similarly, a judge may simultaneously penalize and benefit two different athletes of the same nationality. In those cases, the aggregation of data is likely to lead to the conclusion that bias is not statistically significant because of cancellation effects.

"semi #" refers to the grade awarded by the associated judge, and the subsequent Column with the label "style #" refers to the expected grade associated with that judge's grading style. Grading style is measured following the method in Section 3. Since we have no available information about the judges' grading history before the Olympics, the judges' grading styles are calculated by averaging the grades awarded by each judge during the full Olympic event.

The row with the label "AVERAGE" shows the panel mean and the mean of the grading styles, respectively. The row with the label "% DEV. from AVERAGE" shows the percentage by which the judge Facheng Wang graded the divers differently from the panel mean. The row with the label "% DEV. from STYLE" shows the percentage by which the judge Facheng Wang graded the divers differently from his grading style.

JUDGE	semi 1	style 1	semi 2	style 2	semi 3	style 3
Dennis Gear (NZL)	8.00	7.89	8.00	7.75	8.50	8.46
Facheng Wang (CHN)	8.50	8.05	8.00	7.91	8.50	8.64
Walter Alt (GER)	7.50	7.84	7.00	7.69	8.00	8.41
Bente Johnson (NOR)	7.50	7.88	8.50	7.73	9.00	8.45
Michel Boussard (FRA)	8.00	7.74	8.00	7.60	8.00	8.30
Steve McFarland (USA)	7.50	7.88	7.50	7.74	8.50	8.46
Felix Calderon (PUR)	8.00	7.78	7.00	7.64	8.50	8.34
AVERAGE	7.86	7.87	7.71	7.72	8.43	8.44
% DEV. from AVERAGE	+8.2%		+3.7%		+0.8%	
% DEV. from STYLE	+5.5%		+1.2%		-0.2%	

Table 2: **The men's 3-meter springboard diving competition:** the grades awarded to **Xiong Ni (CHN)** and the expected grades compatible with each judge style in the semi-finals first three dives. Source: Author's own elaboration with data from Emerson et al. (2009).

It is clear from the information in Table 2 that in all three dives performed by Xiong Ni (CHN), the judge Facheng Wang deviated more from the panel mean (i.e., deviations of 8.2%, 3.7% and 0.8%, respectively) than from his own grading style (i.e., deviations of 5.5%, 1.2% and -0.2%, respectively). Similarly, it is clear from the information in Table 3 that in all three dives performed by Fernando Platas (MEX), the judge Facheng Wang deviated less from the panel mean (i.e., deviations of -7.1%, -4.8% and -1.0%, respectively) than from his own grading style (i.e.,

deviations of -9.8%, -7.4% and -3.5%, respectively). This information seems to support the idea that the judge Facheng Wang may have simultaneously benefited the diver Xiong Ni and penalized the diver Fernando Platas.

Both types of deviations are captured by the score aggregation procedure proposed in this paper. However, when applying the aggregation procedure in this paper (see Table 1) the grades of the judge Facheng Wang appear not to be significantly biased and not determinant. The reason might be that the scores awarded in the first dive of Xiong Ni (see Table 2), and the first and second dives of Fernando Platas (see Table 3), were removed from the calculation of the final score. These are apparently the most biased grades. The other three dives, which are considered in the calculation of the final score, are much milder and for that reason not strong enough to induce a significant change in the medal's final standing. This fact may explain why the scores awarded by the judge Facheng Wang seem to have no clear and significant influence on the medal's final standing according to the score aggregation procedure.

JUDGE	semi 1	style 1	semi 2	style 2	semi 3	style 3
Dennis Gear (NZL)	9.00	8.61	8.00	7.89	8.00	7.60
Facheng Wang (CHN)	8.00	8.79	7.50	8.05	7.50	7.76
Walter Alt (GER)	8.00	8.55	8.00	7.84	7.00	7.55
Bente Johnson (NOR)	9.00	8.59	8.00	7.88	7.50	7.59
Michel Boussard (FRA)	9.00	8.44	8.00	7.74	7.00	7.46
Steve McFarland (USA)	8.50	8.60	7.50	7.88	8.00	7.60
Felix Calderon (PUR)	8.50	8.48	8.00	7.78	8.00	7.50
AVERAGE	8.57	8.58	7.86	7.87	7.57	7.58
% DEV. from AVERAGE	-7.1%		-4.8%		-1.0%	
% DEV. from STYLE	-9.8%		-7.4%		-3.5%	

Table 3: **The men's 3-meter springboard diving competition:** the grades awarded to **Fernando Platas (MEX)** and the expected grades compatible with each judge style in the semi-finals first three dives. Source: Author's own elaboration with data from Emerson et al. (2009).

For this reason, i.e., after removing these three extreme scores, in overall terms, the judge Facheng Wang seems to be deviating more from the mean than from his own grading style. Consequently, we can still observe a reversion in the medal's final standing if we place more importance on score deviations from the panel mean (i.e.,

for $\alpha < 0.12$, not shown in Table 1), but not otherwise (i.e., for $\alpha \geqslant 0.12$, as shown in Columns (2) and (3) of Table 1). However, the magnitude of the difference between Fernando Platas and Ni Xiong, even in the most extreme case of $\alpha = 0$, would be very small (i.e., 0.22 points), and for that reason not strong enough to unequivocally support the argument that the judge Facheng Wang has favored Ni Xiong.

These conclusions could have changed drastically, and the score aggregation procedure would have shown more significant corrections, if we have not removed the most extreme grades, as is done by the International Olympic Committee.

Note that the diver ranked in fourth place in Column (1) of Table 1, Xiao Hailiang, is also from China. Table 4 shows the grades awarded and the associated expected grading style of the seven judges in the three semi-final dives in which the judge Facheng Wang participated. The labels and interpretation given to the information in Table 4 are similar to the ones in Table 2. The same is true in the interpretation of the results, where again; the data seem to suggest that the judge Facheng Wang has awarded higher scores to the diver from the same country Xiao Hailiang. The same scoring pattern observed in Table 2 for the diver Ni Xiong is also present in Table 4 for the diver Xiao Hailiang. In other words, in the three dives, the judge Facheng Wang has simultaneously deviated from the overall mean (i.e., 6.8%, 3.8% and 4.4%, respectively) and from his own grading style (i.e., 6.0%, 2.4% and 2.4%, respectively). In the three dives, the judge Facheng Wang was always among the judges awarding the highest score to the diver Xiao Hailiang.

However, this case was not so controversial because the distance between the diver Xiao Hailiang and the diver Dmitri Sautin (ranked in third place) is very large.

In line with our comments, the Olympic Committee and the score aggregation procedure deliver similar numbers in terms of magnitude (see Table 1), which is not necessarily undesirable, since in most cases, bias is only a possibility. Therefore, the score aggregation procedure should correct potential bias, but without distorting the results. In this context, the proposed score aggregation procedure is a refinement of the procedure employed by the Olympic Committee, but it does not dispense with the use of transparency policies like the public disclosure of each judge's grade, which are simply and particularly effective anti-bias monitoring schemes.

5.2. ***The women's 10-meter platform diving competition***

Similarly, Emerson et al. (2009) have also studied the women's 10-meter platform diving competition. They found that judging bias (not necessarily nationalistic bias) could have changed the medals final standing. The diver Laura Wilkinson of USA finished ahead of Li Na of China by 1.74 points (i.e., 543.75 and 542.01 points,

JUDGE	semi 1	style 1	semi 2	style 2	semi 3	style 3
Dennis Gear (NZL)	8.50	8.46	8.50	8.25	8.00	8.18
Facheng Wang (CHN)	9.00	8.64	8.50	8.42	8.50	8.35
Walter Alt (GER)	8.00	8.41	8.00	8.20	8.00	8.12
Bente Johnson (NOR)	8.50	8.45	8.50	8.24	8.50	8.16
Michel Boussard (FRA)	8.50	8.30	8.00	8.09	8.50	8.02
Steve McFarland (USA)	8.50	8.46	8.50	8.24	8.00	8.17
Felix Calderon (PUR)	8.00	8.34	7.50	8.13	7.50	8.06
AVERAGE	8.43	8.44	8.21	8.22	8.14	8.15
% DEV. from AVERAGE	+6.8%		+3.8%		+4.4%	
% DEV. from STYLE	+6.0%		+2.4%		+2.4%	

Table 4: **The men's 3-meter springboard diving competition:** the grades awarded to **Xiao Hailiang (CHN)** and the expected grades compatible with each judge style in the semi-finals first three dives. Source: Author's own elaboration with data from Emerson et al. (2009).

respectively), but after removing the effect of bias, they found that the diver Li Na would have won the event by a margin of 0.36 points. Most of the controversy is driven by the fact that Li Na was well-above Laura Wilkinson after the four semi-final dives, but lost the event in the five final dives. Since both divers finished very close to each other, any potential bias could have made the difference.

However, the identity and the type of bias reported in Emerson et al. (2009) is not clearly specified. Nonetheless, since the score aggregation procedure is constructed to correct for bias, we have applied it to the scores awarded to the nine dives counting to the final score of the women's 10-meter platform diving competition. We found that the application of the proposed score aggregation procedure for $\alpha = 1/3$ and $\alpha = 2/3$ (with $\gamma = 2$ constant), confirms the first place for Laura Wilkinson with 544.68 and 544.61 points, respectively, against Li Na with 541.60 and 541.65 points, respectively. Laura Wilkinson's advantage is reinforced by the score aggregation procedure.

Note that our results do not contradict the results found by Emerson et al. (2009) in support of the existence of bias in favor of Laura Wilkinson. In particular, as pointed out by Emerson et al. (2009) there might exist multiple sources of biases of unknown magnitude affecting both divers in different ways. The difference between our results and their results is justified by the fact that the score aggregation procedure in this paper addresses general forms of bias that are based on deviations from

the panel mean and the judges grading styles. It has not been specifically designed to address nationalistic bias, but it corrects nationalistic bias that materializes either in the form of deviations from the panel mean or the judges' grading style. In this context, our results may suggest that both divers' scores of both divers might have been affected by different forms of bias. Separating and distinguishing between these different cognitive biases is difficult.

6. Conclusion

The existence of bias distorts the quality, reliability, validity and objectivity of the evaluation process, and leads to ineffective decision-making. This issue is relevant in numerous scientific, academic and professional fields.

This paper proposes a practical score aggregation procedure that attempts to reduce and mitigate the influence of bias in subjective judgments. The starting point is to acknowledge that it is virtually impossible to design a procedure that can prevent all forms of bias (Gibbard (1973) and Satterthwaite (1975)). The reason is that conscious bias can be hidden in very complex and strategic ways, and judges are rational agents who can learn how the procedure functions and adjust strategically in order to make bias detection difficult. Consequently, bias is unlikely to disappear, but its influence can be seriously restricted if we adopt adequate bias correction mechanisms.

In this context, the proposed score aggregation procedure offers a tool that can help correct and mitigate the effects of bias that are based on deviations from the panel mean and the judges' grading style. The argument is that biased behavior is associated with either deviations from the mean judgment and/or deviations from the individual judgment style. However, the proposed aggregation procedure has some limitations when it comes to dealing with bias that affects all or the majority of the judges, or bias towards the mean, instead of away from it, and it is not designed to address a particular and specific form of bias (e.g., nationalistic bias), which must be treated individually. For that reason, the proposed procedure does not dispense with the complementary and simultaneous use of transparency policies, such as for instance the public disclosure of each judge's score, which are simple and particularly powerful anti-bias mechanisms (Zitzewitz, 2014).[12] However, in reality,

[12]In this context, in order to introduce bias, judges are forced to award more extreme scores than when the aggregation function is simply the arithmetic mean. Such extreme behavior exposes them to public opinion and to detection by third party monitoring. For this reason, the simultaneous use of transparency policies is important.

and in order to avoid speculation, detailed data about the scores awarded by each judge are usually not publicly available, which creates difficulties when it comes to identifying potential biased behaviors.[13]

In this paper, we focus mostly on sports, but the number of situations that require individual judgments and evaluations, and that can be the object of different sources of bias is endless. The approach in this paper can be extended to these other dimensions of our lives. Nowadays, the internet is making evaluation procedures based on subjective judgments extremely common. Many websites and mobile phone apps ask their users to rate anonymously (or not) all kinds of items, goods and services—from tourist places and blog comments to wines, books, films or music. It is this increasing interest in the content of subjective judgments and their associated controversies that motivates the present paper and the need to study bias in subjective judgments in more detail (Frey and Gallus, 2017; Frey, 2017).

Despite the difficulties associated with the fact that data is not publicly available, and the challenges associated with the design of mechanisms that can prevent or mitigate the influence of all forms of bias, there is plenty of research to be done in this area. A large body of empirical and experimental literature identifies the existence of multiple forms and sources of bias (Bar-Eli et al., 2011; Dohmen and Sauermann, 2016; Plessner and Haar, 2006). However, in most cases, there are no practical or operational solutions that can be applied in real life situations to remove or minimize the negative effects of bias on peoples' lives. This paper is a step forward in this direction and the continuation of an extensive research agenda in bias correction mechanisms in subjective evaluations and judgments.

In this context, we hope this paper will help researchers, practitioners and professionals to better understand how bias operates in subjective judgments, and consequently to provide guidance in the design and implementation of optimal aggregation procedures that can reduce and mitigate the effects of bias in our lives.

Acknowledgments: Financial support from the GRODE Universitat Rovira i Virgili and Generalitat de Catalunya under Projects 2018PFR-URV-B2-53 and 2017SGR770, and the Spanish Ministry of Science and Innovation Project RTI2018-094733-B-100 (AEI/FEDER, UE) is gratefully acknowledged. I would like to thank Jonathan Baron, Juan Pablo Rincón-Zapatero, the Editor and two

[13] In some cases, scores are dissociated from the identity of the judges, which makes bias analysis extremely difficult for researchers and the general public. In other cases (e.g., online judgment of items, goods or services), the data is proprietary and not freely available.

anonymous Referees, as well as several seminars and congress participants for helpful comments and discussions. The usual caveat applies.

References

Arrow, K. J., 1950. A difficulty in the concept of social welfare. The Journal of Political Economy 58 (4), 328–346.

Asch, S., 1951. Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (ed.), Groups, Leadership, and Men, 222–236.

Ashenfelter, O., Quandt, R., 1999. Analyzing a wine tasting statistically. Chance 12 (3), 16–20.

Baker, G. P., 1992. Incentive contracts and performance measurement. Journal of Political Economy 100 (3), 598–614.

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., 2010. Majority judgment: measuring, ranking, and electing. MIT press.

Balinski, M., Laraki, R., 2014. Judge: Don't vote! Operations Research 62 (3), 483–511.

Bar-Eli, M., Plessner, H., Raab, M., 2011. Judgment, decision-making and success in sport. John Wiley & Sons.

Baron, J., 2007. Thinking and deciding (4ed). Cambridge University Press.

Bassett Jr, G. W., Persky, J., 1994. Rating skating. Journal of the American Statistical Association 89 (427), 1075–1079.

Beliakov, G., Pradera, A., Calvo, T., 2007. Aggregation functions: a guide for practitioners. Vol. 221. Springer.

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.

Coupe, T., Gergaud, O., Noury, A., 2018. Biases and strategic behaviour in performance evaluation: The case of the fifa's best soccer player award. Oxford Bulletin of Economics and Statistics 80 (2), 358–379.

Cust, E. E., Sweeting, A. J., Ball, K., Robertson, S., 2019. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. Journal of Sports Sciences 37 (5), 568–600.

Damisch, L., Mussweiler, T., Plessner, H., 2006. Olympic medals as fruits of comparison? assimilation and contrast in sequential performance judgments. Journal of Experimental Psychology: Applied 12 (3), 166–178.

Dawes, R. M., Faust, D., Meehl, P. E., 1989. Clinical versus actuarial judgment. Science 243 (4899), 1668–1674.

Deutsch, M., Gerard, H. B., 1955. A study of normative and informational social influences upon individual judgment. The Journal of Abnormal and Social Psychology 51 (3), 629–636.

Díaz-Pereira, M. P., Gomez-Conde, I., Escalona, M., Olivieri, D. N., 2014. Automatic recognition and scoring of olympic rhythmic gymnastic movements. Human Movement Science 34, 63–80.

Dohmen, T., Sauermann, J., 2016. Referee bias. Journal of Economic Surveys 30 (4), 679–695.

Duggan, M., Levitt, S. D., 2002. Winning isn't everything: Corruption in sumo wrestling. The American Economic Review 92 (5), 1594–1605.

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., 2004. A reputation bias in figure skating judging. Journal of Sport and Exercise Psychology 26 (1), 154–166.

Frank, M. G., Gilovich, T., 1988. The dark side of self-and social perception: black uniforms and aggression in professional sports. Journal of Personality and Social Psychology 54 (1), 74–85.

Frey, B., 2017. Omnimetrics and awards. Tech. rep., Center for Research in Economics, Management and the Arts (CREMA).

Frey, B. S., Gallus, J., 2017. Towards an economics of awards. Journal of Economic Surveys 31 (1), 190–200.

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.

Garicano, L., Palacios-Huerta, I., Prendergast, C., 2005. Favoritism under social pressure. The Review of Economics and Statistics 87 (2), 208–216.

Gibbard, A., 1973. Manipulation of voting schemes: a general result. Econometrica 41 (4), 587–601.

Ginsburgh, V., Noury, A. G., 2008. The eurovision song contest. is voting political or cultural? European Journal of Political Economy 24 (1), 41–52.

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.

Hamermesh, D. S., Schmidt, P., 2003. The determinants of econometric society fellows elections. Econometrica 71 (1), 399–407.

Helsen, W., Gilis, B., Weston, M., 2006. Errors in judging offside in association football: Test of the optical error versus the perceptual flash-lag hypothesis. Journal of Sports Sciences 24 (5), 521–528.

Hilbert, M., 2012. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. Psychological Bulletin 138 (2), 211–237.

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., 1972. Subjective probability: A judgment of representativeness. Cognitive psychology 3 (3), 430–454.

Kahneman, D., Tversky, A., 1996. On the reality of cognitive illusions. Psychological Review 103 (3), 582–591.

Keynes, J. M., 1936. The general theory of employment, interest and money. Kessinger Publishing.

Larsen, T., Price, J., Wolfers, J., 2008. Racial bias in the nba: Implications in betting markets. Journal of Quantitative Analysis in Sports 4 (2), 1–21.

Lee, J., 2008. Outlier aversion in subjective evaluation: Evidence from world figure skating championships. Journal of Sports Economics 9 (2), 141–159.

Lock, R., Lock, J., 2003. The statistical sports fan: judging figure skating judges. STATS 36, 20–24.

Looney, M. A., 2004. Evaluating judge performance in sport. Journal of Applied Measurement 5 (1), 31–47.

Meehl, P. E., 1954. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. University of Minnesota Press.

Nevill, A. M., Newell, S. M., Gale, S., 1996. Factors associated with home advantage in english and scottish soccer matches. Journal of Sports Sciences 14 (2), 181–186.

Osório, A., 2017. Judgement and ranking: living with hidden bias. Annals of Operations Research 253 (1), 501–518.

Page, L., Page, K., 2007. The second leg home advantage: Evidence from european football cup competitions. Journal of Sports Sciences 25 (14), 1547–1556.

Parsons, C. A., Sulaeman, J., Yates, M. C., Hamermesh, D. S., 2011. Strike three: Discrimination, incentives, and evaluation. The American Economic Review 101 (4), 1410–1435.

Pfister, H.-R., Böhm, G., 2008. The multiplicity of emotions: A framework of emotional functions in decision making. Judgment and Decision Making 3 (1), 5–17.

Plessner, H., Haar, T., 2006. Sports performance judgments from a social cognitive perspective. Psychology of Sport and Exercise 7 (6), 555–575.

Popović, R., 2000. International bias detected in judging rhythmic gymnastics competition at sydney-2000 olympic games. Physical Education and Sport 1 (7), 1–13.

Price, J., Remer, M., Stone, D. F., 2012. Subperfect game: Profitable biases of nba referees. Journal of Economics & Management Strategy 21 (1), 271–300.

Price, J., Wolfers, J., 2010. Racial discrimination among nba referees. The Quarterly Journal of Economics 125 (4), 1859–1887.

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.

Shah, A. K., Oppenheimer, D. M., 2008. Heuristics made easy: an effort-reduction framework. Psychological Bulletin 134 (2), 207.

Simon, H., 1955. A behavioral model of rational choice. The Quarterly Journal of Economics 69 (1), 99–118.

Sutter, M., Kocher, M. G., 2004. Favoritism of agents–the case of referees' home bias. Journal of Economic Psychology 25 (4), 461–469.

Tsay, C.-J., 2013. Sight over sound in the judgment of music performance. Proceedings of the National Academy of Sciences 110 (36), 14580–14585.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. Science 185 (4157), 1124–1131.

Unkelbach, C., Memmert, D., 2010. Crowd noise as a cue in referee decisions contributes to the home advantage. Journal of Sport and Exercise Psychology 32 (4), 483–498.

Wang, X. T., Simons, F., Brédart, S., 2001. Social cues and verbal framing in risky choice. Journal of Behavioral Decision Making 14 (1), 1–15.

Wolfers, J., 2006. Point shaving: Corruption in ncaa basketball. The American economic review 96 (2), 279–283.

Wu, S. S., Yang, M. C. K., 2004. Evaluation of the current decision rule in figure skating and possible improvements. The American Statistician 58 (1), 46–54.

Zitzewitz, E., 2006. Nationalism in winter sports judging and its lessons for organizational decision making. Journal of Economics & Management Strategy 15 (1), 67–99.

Zitzewitz, E., 2014. Does transparency reduce favoritism and corruption? evidence from the reform of figure skating judging. Journal of Sports Economics 15 (1), 3–30.