Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making

Eric Zitzewitz[1]

Preliminary

October 2002

[1]Stanford Graduate School of Business, 518 Memorial Way, Stanford, CA 94305. Tel: (650) 724-1860. Fax: (650) 725-9932. Email: ericz@stanford.edu. The author would like to thank Severin Borenstein, Robert Gibbons, Benjamin Hermalin, Phil Leslie, John Morgan, Canice Prendergast, John Roberts, Alan Sorenson, Brian Viard, Justin Wolfers, and seminar participants at Berkeley and George Washington for helpful suggestions and comments.

Abstract

This paper exploits nationalistic biases in Olympic winter sport judging to study the problem of designing a decision making process that uses the input of potentially biased agents. Judges score athletes from their own countries higher than other judges do, and they appear to vary their biases strategically in response to the stakes, the scrutiny given the event, and the degree of subjectiveness of the performance aspect being scored. Ski jumping judges display a taste for fairness in that they compensate for the nationalistic biases of other panel members, while figure skating judges appear to engage in vote trading and bloc judging. Career concerns create incentives for judges: biased judges are less likely to be chosen to judge the Olympics in ski jumping but more likely in figure skating; this is consistent with judges being chosen centrally in ski jumping and by national federations in figure skating. The sports truncate extreme scores to different degrees; both ski jumping and, especially, figure skating are shown to truncate too aggressively; this may contribute to the vote trading in figure skating. These findings have implications for both the current proposals for reforming the judging of figure skating and for designing decision making in organizations more generally.

1 Introduction

Organizations rountinely make decisions for which they need to rely on the informed, but potentially biased opinions of their members. For example, in deciding whether to promote a certain individual or undertake a certain project, the managers who know the individual or project best are often those most likely to be biased. In deciding how much to count the opinions of those who are closely involved, organizations face a trade-off between information and bias. A common solution to this problem is to involve more than one person in the decision. In doing so, organizations face a complicated design problem: how many people to involve, how to aggregate opinions when they differ, how to treat extreme opinions, whether and how to include the opinions of interested parties, whether to strive for continuity in the membership of committees that make similar decisions, and whether and how to adjust the opinions of members based on their histories.

This paper attempts to inform theoretical analysis of this complicated organizational design problem by studying how biased decision makers behave in a team decision making setting. In order to do meaningful empirical work about biases that one might expect to be fairly subtle, we need a large dataset of comparable decisions where individual opinions can be observed and quantified and the expected biases of decision makers can be readily observed by the researcher. As one might imagine, this proved very difficult to obtain in a business setting, so instead we examine an analogous setting in sports: the judging of winter Olympic sports.

Olympic judges represent a particular country and, in the sports we study, they display biases in favor of athletes from the same country. These nationalistic biases can be quite large. Using individual judges' scoring data from the 2002 Olympics and other major international competitions, we find that both figure skating and ski jumping judges score their compatriots about 0.13 standard deviations higher than other judges. In figure skating, where placement is determined entirely by subjective judging, this bias translates to an average placement 0.7 positions higher. These biases are appear even larger when compared to the standard deviation of scores awarded a particular performance; they are about 45 percent of the within-performance standard deviation of scores in both sports. Nationalistic biases are smaller but still positive in mogul skiing, aerials, and the snowboarding halfpipe competition.

In most settings, attempts to study favoritism empirically would be frustrated by the difficulty of observing where we should expect favoritism (e.g., who is "friends" with whom). In this study, judges are biased nationalistically, and thus in a way that we can observe. This allows us to study how the degree of favoritism varies with strategic considerations and how organizational design can be effective or ineffective in dealing with favoritism. We can examine how biases vary with the stakes and with the subjectivity of the aspect being judged. We can examine how judges react to other judge's biases: do they attempt to compensate for each other's biases, or do they form reciprocal arrangements to reinforce each other's biases? In principal, one might expect reputational or career concerns to restrain judges' biases, and we can test whether this is the case. Finally, we can examine whether the methods used to aggregate judges' opinions are appropriate: whether they make optimal use of the information in opinions and whether they create incentives for judges that are conducive to achieving accurate results.

Since both the magnitude of nationalistic biases and our sample size are larger in ski jumping and figure skating than in the other sports, we focus our analysis on these sports. Despite the similarly sized nationalistic biases, in many other respects, ski jumping achieves better results. Ski jumping judges compensate for each other's biases, and do so enough that the net effect on an athlete's total score of having a compatriot on the judging panel is actually slightly negative. In figure skating, however, the compensating biases are positive, so an athlete is much better off when she is represented on the panel. Since judging panels tend to always include the countries that produce the most Olympic skaters, these judging biases represent a sizeable barrier to entry to other countries. The size of nationalistic biases varies significantly by country and judge in both sports. In ski jumping, judges compensate more against athletes from countries represented by a more nationalistically biased judge. The apparent strategic variation of biases depending on the composition of the judging panel is inconsistent with nationalistic biases being the result of only non-strategic differences in tastes; tastes of judges for a certain national style are often given by the ISU as a reason for apparent nationalistic biases in figure skating scores.

The compensating biases in ski jumping, and the fact that they increase when there is more nationalism to be compensated for, are consistent with a desire of judges to maintain the fairness of results in the face of nationalism. In contrast, the positive compensating

biases in figure skating seem less consistent with a desire for fairness than with reciprocity or vote trading. Further evidence of vote trading emerges from the fact that there appear to be long-lived reciprocal relationships between countries. The bias of a judge from country A in favor of an athlete from country B is positively correlated with the bias of a judge from country B in favor of an athlete from country A. The data suggest that the countries most often represented on judging panels can be divided into two blocks, with the U.S., Canada, Italy, and Germany in one block and Russia, the Ukraine, France, and Poland in another block. The ski jumping data rejects similar tests for vote trading. The composition of the voting blocks and the roughly one position size of the bias might seem familiar to those who followed the controversy surrounding the top two finishers in the 2002 Olympic Pairs competition. In fact, these results suggest that the only thing about the judging in the pairs competition that was unusual was that people noticed.[1]

So despite the fact that judges in both ski jumping and figure skating appear to be nationalistically biased, nationalistic biases in ski jumping appear to nearly cancel each other out and thus have a minimal effect on the results, while nationalistic biases in figure skating are magnified by vote trading and long-term coalitions. Although ski jumping and figure skating differ in many ways that might produce these differences, the sports do have institutional features that theory suggests might be contributing factors. Ski jumping and figure skating have different methods for selecting judges for events such as the Olympics. Ski jumping judges are selected by a centralized body (the judges sub-committee of the International Ski Federation [FIS]) while figure skating judges are selected by national federations. Probit regressions predicting selection to judge in the Olympics find that judges with a greater past nationalistic bias are less likely to be selected in ski-jumping but more likely to be selected in figure skating.

Consistent with the importance of reputational concerns, biases rise with the stakes in ski jumping. Biases are largest in the Olympics and near the top of the standings, and are larger in final rounds than in qualifying rounds. At the same time, nationalistic biases

[1] Furthermore, the results in Table 5 imply that with figure skating, nationalistic biases are actually the smallest for Pairs competitions, for major competitions such as the Olympics, and at the top of the standings. While the reports about the role of the Russian mafia in influencing the placement of the top skaters in the 2002 Olympics may lead one to believe that the judging problems in the Olympics were a once-off scandal, the results in this paper suggest that the problems are much more widespread.

exist even for the scored warmup jumps of the top-seeded jumpers who are pre-qualified for the finals; this is despite the fact that these jumps have no effect on the results of the competition. In figure skating, in contrast, biases actually decline slightly with the stakes; they are larger for the compulsories, near the bottom of the standings, and in junior events. These results suggest a tension between the increased extent to which judges care about the results of major events on the one hand and the increase scrutiny given the judging of those events on the other.

Another difference across the sports is the aggregation of scores, specifically, the treatment of extreme scores. In ski jumping, five judges score each jump on style, and the athlete's style points score is the sum of the middle three scores. This method involves some truncation of extreme scores, but the likelihood that a particular judge's score will be influential remains high. In figure skating, the relative ranking of any pair of entrants is determined by which entrant is ranked higher by a majority of judges. This represents the most aggressive truncation of extreme scores possible. Although truncation of extreme scores would be justified if these scores were more likely to reflect observational error or bias than information about the quality of the performance, we present empirical evidence that information content remains high for even those scores. In particular the information-to-noise and information-to-bias ratios do not begin to decline until opinions become $2-3$ times more extreme than the point at which they are normally truncated in ski jumping and about $8-10$ times more extreme than the normal truncation point in figure skating.

Aggressive truncation not only results in less informed judgements, but it can also create incentives for vote trading that lead to further information loss. The intuition is as follows. If a judge from country A wants to influence the results in favor of a compatriot, under aggressive truncation, she can only do it through her own vote to a certain degree. If she cares much more about the ranking of her own-country athletes than she does about the ranking of country-B athletes, and a judge from country B has opposite preferences, the two judges can benefit from vote trading. Assuming that vote trading is non-contractable, however, it must be sustained through repeated interaction. Aggressively truncating opinions by turning them into votes can create a bright line between cooperating with and defecting against a vote trading scheme, since it is easier to determine whether a judge voted for a particular athlete than whether they added 0.2 to the score they would have otherwise given. This

can not only make vote trading easier to sustain, it can also give judges less flexibility to deviate from the agreement in response to the quality of the performances without upsetting the whole agreement.[2] If many observers leave figure skating competitions suspecting that order of finishing was been fixed in advance of the contests, this may be part of the reason why.[3] Although extreme truncation in figure skating is usually justified as a response to judging biases, this paper outlines reasons to believe that it may be making things worse.

The results presented in this paper are potentially interesting at two levels. First, in light of the number of people who watch the Olympics and their non-trivial economic size, the fairness of judging at the Olympics and other sporting events is presumably of direct interest. Second, to the extent that organizations make committee decisions that are similar to Olympic judging and to the extent that institutional features such as the selection of judges and the aggregation of opinions affect the quality of these decisions, the behavior of Olympic judges may have implications for organizational design.

Several recent papers have studied related issues. Recent papers on bias or collusion in sports have found evidence that soccer referees increase the amount of injury time when the home team is behind (Garicano, Palacios, and Prendergast, 2001), that sumo wrestlers throw matches to each other in response to non-linearities in incentives (Duggan and Levitt, 2000), and of point shaving in college basketball games (Wolfers, 2002). Campbell and Galbraith (1996) have measured the nationalistic bias in past Olympic figure skating competitions and found nationalistic biases of similar magnitude to what we find. Others have used data from sports to test theories about behavior in business settings: for example, Bronars and Oettinger (2001) examine the effect of incentives for risk taking in golf tournaments, Goff, McCormick, and Tollison (2002) examine racial integration in baseball to determine whether leaders or followers are more likely to innovate, and Romer (2002) examines the

[2] In some cases, colluding judges attempt to get around the simultaneity problem by communicating in real-time about whether their vote trading agreement still applies given the quality of the performance. For example, on 4/28/02, *60 Minutes* broadcast a tape of two judges communicating with glances and foot signals. The difficulties of communicating in real-time, however, particularly with cameras rolling, make it very difficult to make collusive agreements contingent on performance quality.

[3] For example, prior to the 1998 Olympic Ice Dancing competition, Ukrainian judge Yuri Balkov was taped by another judge announcing the order in which he would rank contestants. As further evidence that career concerns do not create strong incentives for fairness in figure skating, Balkov was temporarily suspended by the International Skating Union but was selected by Ukraine to judge at the 2002 Olympics.

decision to punt on fourth down in American football as a dynamic programming problem.

There is also an extensive theoretical literature on the problems of relying on information from potentially interested parties. The simple model in the next section takes an approach that is closest to that of Prendergast and Topel (1996), who examine the problem of relying on the opinion about employee performance from one potentially biased supervisor.[4] Aghion and Tirole (1997) and Athey and Roberts (2001) examine the related problem of trusting the opinion of an employee about the quality of a project from which she will derive some private benefit. Milgrom and Roberts (1986) examine the situation where employees cannot falsify but can hide information, and find that under certain circumstances, having one advocate on either side of an issue yields full revelation of information. Finally, there is an extensive recent theoretical and empirical literature on the career concerns of forecasters or opinion producers.[5] While this literature primarily focuses on the career concern of appearing to have high-quality signals, our empirical evidence suggests that the career concerns of appearing unbiased (or, in the case of figure skating, biased) is more important for Olympic judges.

The remainder of the paper is divided into four sections. The first section considers the problem of a principal who is attempting to design a mechanism for aggregating the opinions of potentially biased judges, and informally discusses the effects of reputational concerns or a taste for fairness. The second section contains the empirical results summarized above. Two concluding sections follow. One discusses the proposals for reform of figure skating judging made after the 2002 Olympics, and the extent to which they appear sensible in light of the results of the paper. The second discusses the extension of the results to a corporate setting, in which committees iterate until a consensus or near consensus is reached, rather than aggregating simultaneously announced opinions.

2 Aggregating opinions of biased judges

This section analyzes the mechanism design problem faced by a principal who is attempting to construct an evaluation of a particular performance (or project) quality using the opinions

[4] It is also similar to Meyer, Milgrom, and Roberts (1992) in which an employee who does not want to get fired incurs influence costs to raise her employer's signal of project quality.

[5] CITES

of potentially biased judges (or committee members). Judges observe a noisy signal of the quality of the performance and choose their opinion to balance two conflicting objectives: biasing the result of the evaluation in a particular direction and reporting an accurate opinion to maintain their professional reputation. Performance quality is assumed to be objective or, alternatively, defined according to the tastes of the principal; we consider any difference in taste's with the principal that affects their scores to be part of a judge's bias. Judges announce their opinions simultaneously, and the principal's mechanism aggregates these opinions into an evaluation of the performance.

We assume that judges construct their opinions using a prior belief that is common to all judges and other observers and a private, noisy signal of performance quality. Performance quality is unidimensional and is given by q, and we calibrate performance such that the common prior is that $q \sim N(0,1)$. Each of the J judges observes the prior plus a noisy signal $s_j = q + n_j$, where n_j is the judge's observational error, with $E(n_j|q) = 0$.

Judge j chooses her opinions m_j to solve the following problem:

$$\max_{m_j} E[-\frac{(m_j - q)^2}{2} + b_j \cdot y(\mathbf{m})|s_j], \tag{1}$$

where $y(\mathbf{m})$ is the function used to aggregate the opinions into an evaluation. The variable b_j captures the judge's bias: the extent that she cares about the evaluation $y(\mathbf{m})$ directly, relative to her concern for issuing an accurate opinion. For simplicity, the setup above assumes that judges care directly about absolute evaluations, not the evaluation of one performance relative to another as might be in case in a tournament. Of course, in a two-competitor tournament, we could simply define q to be the difference in performance quality, and m_j and $y(\mathbf{m})$ to be opinions about and evaluations of this difference.

A judge's bias is known to the judge, but not to the principal or other judges. We assume for the moment that judges have no long-run concerns (e.g., reputational concerns) and that they care only about the accuracy of their opinion, i.e. about $(m_j - q)^2$, not the accuracy (or fairness) of the final evaluation, i.e. about $(y - q)^2$. The judge's optimal opinion is:

$$m_j[s_j, b_j, y()] = E(q|s_j) + b_j \cdot E(\frac{\partial y}{\partial m_j}|s_j). \tag{2}$$

Judges report their best estimate of performance quality, plus the product of their bias and the extent to which they expect their opinion to be influential.

The principal's problem is to produce the best possible evaluation; we will assume a quadratic loss function for the principal, so best means minimum mean-squared-error. If she cannot commit to a mechanism and has no long-run concerns, then she just sets $y = E(q|\mathbf{m})$ taking the opinions as given. If she can commit, then she solves the problem

$$\max_{y()} E\{-[y(\mathbf{m}(\mathbf{s}), \mathbf{b}, y()) - q]^2\}. \tag{3}$$

The principal will want to put more weight on opinions that are more likely to be informative: either because they come from a judge who the principal believes to make smaller observational errors and/or be less biased or because the opinion's magnitude causes the principal to expects the signal-to-noise and/or signal-to-bias ratios to be especially high. For example, if extreme opinions are more likely to reflect noise or bias than less extreme opinions, then the principal would want to discount extreme opinions. The problem is, the act of putting more weight on an opinion that is from a certain judge or in a certain range will increase the bias content of that opinion, making the principal at least partly regret putting the extra weight on it.

This problem limits the extent to which a principal will be able to benefit from committing to aggregate scores in a certain way. In Appendix A, we argue that when observational errors and biases are normally distributed, and thus signal-to-bias and signal-to-noise ratios do not decline for extreme opinions, then the optimal aggregation scheme is very close to linear, even when the principal has commitment power. We later find that the actual aggregation mechanisms used by figure skating and ski jumping truncate scores well before the point at which the signal-to-noise and signal-to-bias ratios begin to decline, suggesting that they are deviating from optimality.[6]

[6] In Appendix A, we assume a quadratic loss function for the principal, but the result that she does not want to deviate from linearity should not be very sensitive to the shape of her loss function, so long as the loss function is symmetric, since if prior beliefs, biases, and observation errors are normally distributed, then a principal's posterior belief will be as well, and a principal's expectation of q will also be her modal belief. The result in Appendix A could be different if a principal had perferrences about things other than $y - q$. For example, if the principal cared directly about the appearance of a consensus, as a sport's governing body might, then this would create a new rationale for truncating extreme opinions.

2.1 Effect of reputational concerns and a taste for fairness

Biased judges receive utility from changing the outcome of the mechanism in the direction of their bias. If the principal uses a linear mechanism, with $y = \sum_j w_j \cdot m_j - c$, then this extra utility is equal to the product of the judge's bias b_j, her weight in the mechanism w_j, and the extent to which she will bias her opinion: $b_j w_j$, i.e. a extra utility of $(b_j w_j)^2$. If judges' biases are uncorrelated, then a judge whose w is reduced will lose some of this extra utility, since her weight given her opinion will be replaced by an weight on opinion that is unbiased in expectation.

When a principal has information on a judge's history, then she should be able to improve upon a symmetric mechanism by taking this history into account. The simplest case to analyze is that of a principal without commitment power who sets $y = E(q|\mathbf{m})$. If judge's observational errors and biases are uncorrelated, then the principal will subtract a judge's expected bias and then weight each opinion in proportion to the inverse of its mean squared error:

$$y = \sum_j w_j \cdot (m_j - \widehat{b}_j \cdot w_j),$$

with w_j given by

$$w_j = \frac{\lambda}{w_j^2 \cdot \widehat{Var}(\widehat{b}_j) + \widehat{Var}(n_j)},$$

where λ is some constant and $\widehat{b}_j$ is the principal's expectation of judge j's bias. Judges who will issue future opinions face a trade-off between the utility gain from biasing their opinions today at the cost of increasing the perceived bias that will be backed out of their future opinions. These reputational concerns have effects analogous to those in Holmstrom (1999): "young" judges will surpress their biases in order to retain influence in the future, while judges nearing retirement will bias their opinions more fully.

In addition, judges may have an incentive to issue opinions that are similar to other judges', since deviating from the opinion's of other judges might reduce the principal's faith in her observation ability and increase $\widehat{Var}(n_j)$. For the same reason, judges may have an incentive to be consistent in their appraisal of the performance of a given athlete, to the extent that performance quality is influence by athlete ability about which there is uncertainty.

If a judge has better information than the principal about another judge's biases, then

the incentive to issue opinions that are similar to other judges' can lead a judge to bias her opinion in the direction of other judge's biases (Prendergast, 1993). At the same time, if a judge has a concern about the fairness of the evaluation, as opposed only caring about the accuracy of her own opinion, then we might expect judges to partially undo biases in other judges opinions. Thus if we observe judges reinforcing each other's biases, we might interpret this as resulting from reputational concerns, while if we observe them undoing each other's biases, this might be the result of a taste for fairness.

3 Evidence on nationalistic biases

In this section, we will analyze individual judges scoring of athletic performances in winter sports. While judges can be biased in favor or against an athlete for many reasons, one of the easiest to observe is nationalistic biases. Once we document the rather large nationalistic biases that exist in the data, we will conduct several tests motivated by the theoretical discussion above:

1. *Do nationalistic biases increase with the stakes?* If judges care more about the outcome of events such as the Olympics, we would expect larger biases in scoring.

2. *Do judges strive for consistency?* The highest seeded ski jumpers are pre-qualified for the finals, but are still allowed to take a scored warm-up jump in the qualifying round. If judges are attempting to maintain consistency in their scoring of a particular athlete, we should still observe a bias even for these scores for which the stakes are zero.

3. *Do judges undo or reinforce the nationalistic biases of other judges?* If judges have a taste for fairness, then we should expect them to bias against an athlete from a country that is represented on the judging panel, especially when the judge is question has a history of nationalistic bias. At the same time, the judges may reinforce each other's biases in order to appear less biased themselves, or as part of a collusive arrangement.

4. *Do judges' actual career concerns encourage unbiased judging?* We can observe the correlation between being chosen for the Olympics and a judge's apparent bias in judging earlier events to examine the incentives created by career concerns.

5. *Are extreme scores informative?* Both ski jumping and figure skating use non-linear mechanisms for aggregating scores that underweight extreme scores. The discussion in Section 2 and Appendix A suggests that this would only be optimal if extreme scores were more likely to indicate bias or observational error than information about performance quality. We can test whether this is the case.

3.1 Data

The data are individual judge's scorings of figure skating, ski jumping, mogul skiing, aerials, and snowboarding performances from the 2002 Winter Olympics and other events immediately before or after the Olympics. For figure skating, the sample includes all events for which score sheets containing individual judge's scores were available either on the International Skating Union website or elsewhere on the web.[7] For the other sports, the sample includes the Olympics, all World Cup events, and the World Junior Championships; for these events, score sheets, as opposed to just results, were available on the International Skiing Federation web site.

Table 1 summarizes the dimensionality of the dataset. We have data on 16 figure skating events, 25 ski jumping events, and about 8 events each for moguls, aerials, and snowboarding. All but one figure skating event includes separate competitions for men, women, pairs, and ice dancing, while almost mogul, aerials, and snowboarding events include competitions for both men and women. About half of figure skating competitions include compulsory rounds, and likewise about half of ski jumping, mogul, and aerials contests include qualifying rounds. All figure skating competitions include two performances in addition to any compulsories; in ski jumping the finals include two jumps for each competitor, while in snowboarding and aerials finals the top competitors after the first performance are allowed a second. We have data on close to 3,000 athletic performances in figure skating and ski jumping, but only roughly 1,000 or less in the other sports.

[7] The ISU and other skating organizers use a software program called IceCalc to tabulate results and generate score sheets. Score sheets on websites other than the ISU's were found by conducting a google search for "Created by IceCalc," which is inserted on every score sheet. In addition to the Olympics, the figure skating sample includes the 2001 and 2002 European championships, the 2001 World and World junior championships, the 2001 and 2002 Four Continents competition, the ISU junior championships, and several other events.

Table 2 presents summary statistics for the judges' scores. In figure skating each of 5, 7, or, usually, 9 judges scores each performance on two dimensions, technical merit and artistic impression. Skaters are then ranked ordinally by each judge based on the sum of these scores. We analyze both the judges' ordinal ranking and the sum of the scores, although the results are understandably very similar. In ski jumping, each jump is scored on style by five different judges. In the other three sports, judges are assigned specific aspects of the performance to judge. Especially in moguls and aerials, the scores for different aspects have different means, standard deviations, and ranges, suggesting that they are not strictly comparable. This is important because we identify nationalistic biases partly by comparing the scores given the same performance by different judges, and the extent to which scores are not comparable across the aspects being judged further reduces the amount of data we have to work with in these sports.

The aggregation of scores differs across the sports, particularly in the extent to which extreme scores are truncated or underweighted. In snowboarding there is no truncation; the five scores are summed to yield a total score. In ski jumping, the total style score is the mean of the middle three of five scores. Moguls and aerials use a combination of these two approaches: the mean of the middle three scores for the aspect judged by five judges (turns in moguls, air and form in aerials) is added to the two scores for the other aspect (air in moguls, landing in aerials). In aerials, the score is then multiplied by a degree of difficulty factor for the jump attempted.

Figure skating uses the most extreme form of truncation. The scores given each performance for technical merit (TM) and artistic impression (AI) are added together and then each judge assigns each competitor an ordinal rank, with any ties broken in favor of TM in the short program and AI in the long program. Placement in each round is then determined by majority vote: a competitor is ranked ahead of another if he/she/they place higher with a majority of the judges.[8] Overall placement is then determined by ranking the skaters on a weighted average of their placements in each round, with the short program weighted more than the complusories but less than the long program. In summary, the only aspect of a judge's scores that matters is the relative ranking of two skaters: the difference in the scores

[8] This leads to occasional non-transitive preferrences in which majorities prefer skater A to B, B to C, and C to A. These are resolved in favor of the skater of the three who wins the most bilateral comparisons.

assigned to the two skaters does not affect the results of the competition. The ranking of the sports in terms of the extent to which they truncate extreme scores is therefore: figure skating, ski jumping, moguls/aerials, and snowboarding.

A final note is that while subjective scoring accounts for 100 percent of a performance's score in figure skating, aerials, and snowboarding, in ski jumping and moguls there is also an objective component: distance jumped and time, respectively. In ski jumping, the objective component accounts for a large share of the variance in results: the standard deviation of distance points, style points, and total points are 20.9, 3.3, and 22.9, respectively. In moguls the reverse is true: time points, turns and air points, and total points have standard deviations of: 1.0, 3.1, and 3.8, respectively. In both sports, the covariance between subjective and objective scores is positive.

3.2 Measuring nationalistic biases

Our primary empirical problem in measuring nationalistic biases is that we do not observe an objective measure of performance quality. We can, however, draw inferences about performance quality from other judges' scores, scores given to other performances by the same athlete, and, in the case of ski jumping and mogul skiing, objective measurements of the distance and speed of the jump or the time of the run.

We can write the score given by judge j to performance p by athlete i as:

$$s_{ijp} = q_{ip} + B_{ij} + e_{ijp}, \tag{4}$$

where s_{ijp} is the score, q_{ip} is the objective quality of the performance, B_{ij} is the bias of judge j in favor of athlete i, and e_{ijp} is the judge's observational error.[9] We write B_{ij} for the bias in scores to distinguish it from the bias parameter of the utility function b_j in the prior section; B corresponds to $b \cdot w$ above. As in Section 2, performance quality is considered to be objective, and any influence of a judge's personal tastes on the scores is considered to be bias. We likewise write e_{ijp} for the observational error component of the judge's score;

[9] In figure skating, we take our primary measure of the score given a performance by a judge to be the sum of the technical merit and artistic impression scores. Given the fact that only the ordinal ranking given by each judge matters for the final standings, we also repeated our analysis using the ordinal placement as the dependent variable, and obtained very similar results.

when judges have prior information about performance quality, this will be smaller than the observational error n_j in section 2.

Two strategies for identifying nationalistic biases are to: 1) compare the scores for the same performance given by different judges, or 2) use other observables to infer performance quality and assume that the remaining uncertainty in performance quality is uncorrelated with judge and athlete affiliation. The first approach is the less data-intensive of the two, but it only allows one to measure the difference between judges' bias in favor of their own athletes (nationalistic bias) and in favor of or against athletes from other countries that are represented on the panel (compensating bias). If we are willing to assume that compensating biases are small, then we can view this difference as being an approximation of the nationalistic bias.

Estimating average nationalistic bias using this first approach involves estimating the model:

$$s_{ijp} = B \cdot \Phi(I = J) + q_{ip} + e_{ijp} \tag{5}$$

where B is the average nationalistic bias, I and J index athlete and judge countries, and q_{ip} is a performance fixed effect. Table 3 reports results from this method for the five sports. Statistically significant biases exist in figure skating, ski jumping, and mogul skiing. These biases are arguably also "economically significant," particularly given their size relative to the within-performance standard deviation of scores.

Table 4 uses the second identification approach, which allows us to relax the assumption of zero compensating bias. In focuses on the sports with both a large sample size and a significant nationalistic bias: figure skating, ski jumping, and the turns aspect of mogul skiing. The second approach involves estimating the model:

$$s_{ijp} = B_{nat} \cdot \Phi(I = J) + B_{comp} \cdot \Phi(I \in P, I \neq J) + a_i + \beta x_{ip} + \widetilde{q}_{ip} + e_{ijp} \tag{6}$$

where P is the set of countries represented on the judging panel, a_i is an athlete fixed effect, x_{ip} is a vector of observables about the performance, and $\widetilde{q}_{ip}$ is a performance random effect that captures the variation in performance quality that is unexplained by a_i or x_{ip}. Identification involves assuming that the component of performance quality that is uncorrelated with a_i or x_{ip} is uncorrelated with the composition of the judging panel.

For figure skating, the x_{ip} includes fixed effects for meet*event*round combinations.

Identification thus involves assuming that a particular athlete's performance is not correlated with the composition of the judging panel, except to the extent that all athletes score higher or lower in a given meet*event*round. For moguls and ski jumping, the x_{ip} also includes the time (for moguls) and the distance jumped and takeoff speed for each jump (for ski jumping). All else equal, a mogul skier with better form will descend faster, and a ski jumper with the same takeoff speed but better form while airborne will travel further. For example ski jumping, if we regress style points on distance and speed, we find positive and negative coefficients on distance and speed respectively, with the regression explaining about 50 percent of the variation in style points. If we call the style that is statistically explained by distance and speed "airborne" style and the residual "landing" style, then including jump characteristics relaxes our identification assumption to assuming that a given athlete does not have especially good or bad landings (as opposed to overall performances) when judges from particular countries are on the scoring panel.

The results in Table 4 suggest that measuring $B_{nat} - B_{comp}$ is much easier than separately measuring its components, but we can obtain statistically significant estimate of B_{comp} for figure skating and ski jumping. These estimates imply that there is a negative compensating bias in ski jumping but a positive bias in figure skating. Ski jumping judges undo the nationalistic biases of their colleagues, while figure skating judges reinforce them. The net effect of nationalistic and compensating biases in ski jumping on the total style point score is actually ambiguous: a regression of total style point score on the presence of a compatriot on the panel and skier and meet*round fixed effects yields a point estimate of -0.042 ($SE = 0.15$); a similar regression for the median score in figure skating yields a estimate of 0.116 ($SE = 0.025$), which implies that having a compatriot on the judging panel is very important. The positive B_{comp} in figure skating and negative B_{comp} in ski jumping also implies that the first identification approach overstates nationalistic biases in ski jumping while understating them in figure skating.

3.3 Variations in nationalistic biases

In this subsection we estimate nationalistic biases for subsamples of the data. When examining subsamples, we use the first identification approach and limit the analysis to ski jumping and figure skating due to sample size constraints. Table 5 presents estimates of

nationalistic bias for subsamples of the data.

In ski jumping, nationalistic biases are larger when the stakes are higher, consistent with judges caring more about the outcomes when the stakes are higher. Nationalistic biases are larger: 1) in the Olympics, 2) among the higher placing skiers in the final round, 3) among skiers that have not already pre-qualified in the qualifying round, 4) for team events, and 5) on the 90m hill, where style points account for a large amount of the variance in total scores. Nationalistic biases exist even for the scored qualifying jumps of pre-qualified skiers, which have no effect on the competition.

In figure skating, the pattern is different. Nationalistic biases are smaller when the stakes are higher. Biases are smaller in the Olympics than in other major events, and they are largest in junior competitions. Biases are larger for the short program, which receives a lower weight, than for the long program. Biases are also larger for the complusories, which are weighted less than either the short or long programs. Biases are larger where scoring is more subjective, as it is for ice dancing, where skaters do not have as many mandatory deductions for falls, and for artistic impression as opposed to technical merit scores. Arguably, this is also true in mogul skiing, where biases are larger for turns than for air, since a component of the score for air is the height of the jump, which might be considered more objective than the form of the turns.

A possible explanation of this difference is that whereas the scrutiny of style judging in ski jumping may be limited, in 2001-2 scrutiny was fairly significant in figure skating and was probably especially so for more important competitions and for non-complusory rounds that are watched by larger audiences. In terms of the model in Section 2, b, the ratio of the importance of the event and the reputational costs of appearing biased, may actually be lower in figure skating for higher stakes events.

Another source of variation in nationalistic bias is by country. Table 6 presents estimates of country-specific nationalistic bias estimated using the following model, which is a version of (5):

$$s_{ij} = B_J \cdot \Phi(I = J) + L_J + q_{ip} + e_{ijp}. \tag{7}$$

B_J captures the bias of judges from a particular country in favor of their own athletes. L_J captures the "leniency" of the judge country: the extent to which all scores issued by judges from that country are higher or lower than their counterparts. The q_{ip} are performance

fixed effects as in (5). One might note that nationalistic biases appear to rise with eastern longitude, and they also appear to rise with the various indices of country-level corruption that have been calculated.

If a desire for fairness explains the negative compensating bias in ski jumping, we should observe judges compensating more against athlete from countries with a history of being more nationalistic. Table 7 presents results from estimating a version of (6) that allows B_{comp} to vary with the nationalistic bias of the athlete's representative on the judging panel:

$$s_{ij} = B_{nat} \cdot \Phi(I = J) + (C_{comp} + D_{comp} \cdot \widehat{B}_I) \cdot \Phi(I \in P, I \neq J) + a_i + \beta x_{ip} + \widetilde{q}_{ip} + e_{ijp}, \quad (8)$$

where $D_{comp} < 0$ indicates a desire for fairness and C_{comp} is the "panel representation effect", or the bias in favor of athletes represented by a judge from an unbiased country. In results in Table 7 imply that both ski jumping and figure skating judges are more biased against athletes from countries with more biased judges. The major difference is in the size of the panel representation effect, C_{other}, which in figure skating is positive and large enough that the net compensating bias is still positive on average. In addition, D_{comp} is also slightly larger in ski jumping, indicating that the extra nationalism of a particular judge is more than compensated for in each of the other judges' scores.

The fact that D_{comp} is negative in both sports is inconsistent with career concerns being the explanation for the fact that figure skating judges bias in favor of represented athletes. If judges were concerned with being outliers and therefore thus biased in favor of represented athletes since they knew that the judge representing this athlete would be biased, then we would expect D_{comp} to be positive, i.e. for judges to bias the most in favor of athletes from the most nationalistic countries.

One remaining potential explanation is collusion: judges bias in favor of represented athletes because they expect something in return. To test for collusion, we can test for whether biases are reciprocal. To do this, we estimate a version of (6):

$$s_{ij} = B_{IJ} + a_i + \beta x_{ip} + q_{ip} + e_{ijp}, \quad (9)$$

where B_{IJ} is the average bias of a judge from country J in favor of an athlete from country I. We can then perform a simple test for reciprocity by examining the correlation between $\widehat{B}_{IJ}$ and $\widehat{B}_{JI}$. In figure skating, the $\widehat{B}_{IJ}$ and $\widehat{B}_{JI}$ are positively correlated, with a correlation

coefficient of 0.122 (p-value 0.07) for the top-10 judging countries, while in ski jumping, they are negatively correlated, with a coefficient of −0.125 (p-value 0.24).

Given the suggestions of "bloc judging" in figure skating, we also test for whether bloc voting can help explain the patterns in the $\widehat{B}_{IJ}$. We estimate by maximum likelihood a model in which there are two voting blocs, and $B_{IJ} = B_{same}$ if I and J are members of the same bloc or $B_{IJ} = B_{diff}$ if they are different. We allow the nationalistic biases, i.e. the B_{II}, to vary freely for each country. In figure skating, likelihood is maximized for the top-10 countries by a model in which the U.S., Canada, Germany and Italy are in one bloc, and France, Poland, Russia, and the Ukraine are in another. Japan and China are not consistently classified in one bloc or the other, and thus could be thought of as non-aligned (Table 8).[10] The estimate $\widehat{B}_{same} = 0.001$ and $\widehat{B}_{diff} = -0.051$, so together with Table 2 these results imply that a typical judge biases by 0.17 in favor of her own athletes and by −0.05 against athletes from the other voting bloc. Given the lack of reciprocity in ski-jumping noted above, the data rejects the voting bloc model.

3.4 Career concerns of judges

As we argued in Section 2.1, the desire to maintain influence in the future may motivate judges to moderate their biases. A principal who is attempting to minimize the mean squared error of the evaluations would want to underweight the opinion of judges who have appeared to be biased in the past or who appear to make large observation errors (see equation 4, above).

In sports judging, the only way to underweight a judge's opinion is to exclude her from the judging panel; once the judging panel is chosen, all judges opinions are given equal weight.[11] Olympic judges are unpaid, so presumably their only career concern is maintaining their future influence by being chosen to judge important competitions. Given that our sample ends with the 2002 Olympics, a natural way to test whether Olympic judges face career-concern related incentives to moderate their nationalistic biases is to examine the determinants of being selected to judge in the Olympics.

[10] The results exclude the top 2 couples in the Pairs competition at the 2002 Olympics, in which the media speculated there had been bloc voting along roughly these lines.

[11] This excludes the very rare cases when a judge's opinion is invalidated after the fact, such as when the French judge's opinion was discarded in the 2002 Olympic Pair competition.

Table 6 examines the relationship between a individual judge's performance in the 14-17 pre-Olympic events and whether or not she is chosen to judge in the Olympics. Panel A compares the judges chosen with those not chosen along three dimensions: the leniency of their scoring, their observed nationalistic bias, and the consistency of their scoring with that of the other judges. If we call *d-score* the difference between a particular judge's scoring of a performance and the average score of the other judges, then leniency is defined as the average *d-score* across all observations, nationalism as the difference between the average *d-score* for compatriots and for all athletes, and consistency the average absolute value of *d-score*. Since the selection process for judges is often two-stage, which countries to be represented chosen first and representative chosen second, we compare the chosen judges with both all the judges who were not chosen and with those that were not chosen but represent the same country, but the results are similar in character regardless of the comparison group.

Panel A suggests that the most important distinction of the ski jumping judges who were selected for the Olympics was that they displaced essentially no nationalistic bias.[12] This difference is statistically significant, as is the coefficient in the multivariate probit regression in Panel B. For figure skating, however, the judges chosen for the Olympics are both statistically significantly more lenient and *more* nationalistic. The probit regressions find these two factors to be jointly significant (p-value = 0.02), but neither individually significant.

The finding that nationalistically biased judges are less likely to be chosen in ski jumping, but more likely to be chosen in figure skating, is not surprising as it might first appear given how the judges are actually selected. In ski jumping, judges are selected by a centralized committee, which could potentially act as a principal interested in achieving minimum mean-squared error scoring. In figure skating, judges are nominated by their national federations. Given that federations presumably get considerable utility from seeing their own athletes win, sending a biased judge is presumably privately optimal. Since 20 different national federations send judges to the Olympics, it is quite plausible that cooperation on selected unbiased judges is difficult to sustain.

[12] Although six ski jumping judges judged in the 2002 Olympics, only three judged 2001-2 World Cup events that were in our sample.

3.5 Non-linear aggregation of scores

Both ski jumping and figure skating underweight extreme opinions. In ski jumping, total style points are the sum of the middle three of five scores. If a score is already tied for the highest or lowest, raising or lowering it (respectively) does not affect the total style points awarded. Figure skating is even more extreme: all that matters is which skater ranks higher on a judge's scorecard, by how much does not matter (except when resolving non-transitivities as discussed above).

The discussion in Section 2 and Appendix A implies that the optimal aggregate mechanism should only underweight extreme opinions if the signal-to-noise and signal-to-bias ratios of opinions decline as they become more extreme. The logic behind this is that if signal-to-noise and signal-to-bias ratios remain high, then there is information in extreme opinions, and underweighting them would be discarding this information. Underweighting extreme opinions is often justified in terms of reducing the incentive to bias opinions, but as we discuss above, deviating from linearity has the perverse effect of underweighting exactly those opinions that will be less biased.

How can we determine whether the signal-to-noise and signal-to-bias ratios decline as opinions become more extreme? If we observed an objective measure of performance quality, we could examine the predictive power of extreme opinions. In other words, we could ask: when one judge's opinion is very different from the others, how much weight should we be putting on the extreme opinion in constructing an estimate of performance quality? If we take the mean of the opinions of $J-1$ judges as a starting point, how should we revise our expectation of quality based on the difference between the Jth opinion? We would like to estimate the function $f()$ in:

$$E(q_{ip}|m_{ipj}, \overrightarrow{m}_{ip,-j}) - E(q_{ip}|\overrightarrow{m}_{ip,-j}) = f(m_{ipj} - \overline{m}_{ip,-j}). \tag{10}$$

If we observed q_{ip} and had a method for estimating $\widehat{E}(q_{ip}|\overrightarrow{m}_{ip,-j})$, we could estimate this function by non-parametrically estimating:

$$\begin{aligned} q_{ip} - \widehat{E}(q_{ip}|\overrightarrow{m}_{ip,-j}) &= f(m_{ipj} - \overline{m}_{ip,-j}) + \varepsilon_{ipj}. \qquad (11) \\ \varepsilon_{ipj} &= [q_{ip} - E(q_{ip}|m_{ipj}, \overrightarrow{m}_{ip,-j})] + [E(q_{ip}|\overrightarrow{m}_{ip,-j}) - \widehat{E}(q_{ip}|\overrightarrow{m}_{ip,-j})] \end{aligned}$$

This estimate will be consistent if the expectation of the error term is zero for all values of the right-hand side variable: $E[\varepsilon_{ipj}|m_{ipj} - \overline{m}_{ip,-j}] = 0$. We know this is true for the first

term, the expected error in the expectation $E(q_{ip}|m_{ipj}, \overrightarrow{m}_{ip,-j})$ must be zero conditional on all functions of m_{ipj} and $\overrightarrow{m}_{ip,-j}$. It will be true for the second term so long as we do not construct our estimate of $\widehat{E}(q_{ip}|\overrightarrow{m}_{ip,-j})$ in such a way that its error is correlated with $m_{ipj} - \overline{m}_{ip,-j}$. Up to this point, this methodology is similar to the methodology used for estimating exaggeration and the information content of analyst's earnings forecasts introduced in Zitzewitz (2001) and Zitzewitz (2002).

An important difference between the judging data and the analyst data is that in the judging data we do not observe an objective measure of q_{ip}, so we cannot estimate (11). What we can do instead is take the opinion of judge k to be the objective measure and then study how our expectation of m_{ipk} changes with $f(m_{ipj} - \overline{m}_{ip,-jk})$. This adds a term to the error term: $m_{ipk} - q_{ip} = B_{ik} + e_{ipk}$, the sum of the bias and the observational error, which we abbreviate $x_{ipk} = m_{ipk} - q_{ip}$. If we use $\overline{m}_{ip,-jk}$ as our proxy $\widehat{E}(q_{ip}|\overrightarrow{m}_{ip,-jk})$, then the regression equation and error term become:

$$\begin{aligned} m_{ipk} - \overline{m}_{ip,-jk} &= f(m_{ipj} - \overline{m}_{ip,-jk}) + \varepsilon_{ipj}. \qquad (12) \\ \varepsilon_{ipj} &= [q_{ip} - E(q_{ip}|m_{ipj}, \overrightarrow{m}_{ip,-jk})] + \\ &\quad [E(q_{ip}|\overrightarrow{m}_{ip,-jk}) - \overline{m}_{ip,-jk}] + \\ &\quad m_{ipk} - q_{ip} \end{aligned}$$

As above, the first term must be zero in expectation for all values of $m_{ipj} - \overline{m}_{ip,-jk}$. The third term is x_{ipk} while the second term is equal in expectation to $-\overline{x}_{ip,-jk}$ for all values of $m_{ipj} - \overline{m}_{ip,-jk}$. Since $m_{ipj} - \overline{m}_{ip,-jk} = x_{ipj} - \overline{x}_{ip,-jk}$, our identification assumption becomes: $E(x_{ipk} - \overline{x}_{ip,-jk}|x_{ipj} - \overline{x}_{ip,-jk}) = 0$.

An easy way of ensuring that $x_{ipk} - \overline{x}_{ip,-jk}$ is uncorrelated with $x_{ipj} - \overline{x}_{ip,-jk}$ is to include all combinations of j and k for each observation, which make them uncorrelated by construction. For identification, however, it is necessary that $x_{ipk} - \overline{x}_{ip,-jk}$ not be related to higher moments of the distribution of $x_{ipj} - \overline{x}_{ip,-jk}$. Given our finding that there is a surprising amount of information in extreme opinions, what we should worry most about is $x_{ipk} - \overline{x}_{ip,-jk}$ being positively related to the skew in the $x_{ipj} - \overline{x}_{ip,-jk}$. This might be the case if judges made errors and had biases that were large in absolute value with a small probability *and* if these large errors and the likelihood of making them for a given performance were correlated. After discussing the results, we will discuss their robustness

to this potential problem.

The slope of our estimated $f()$ gives us the incremental signal-to-message ratio. The signal-to-message ratio could be low due to either noise or bias. To get an understanding of the bias-to-message ratio, we also estimate the function $g(m_{ipj} - \overline{m}_{ip,-jk}) = E(B_{IJ}|m_{ipj} - \overline{m}_{ip,-jk})$. This captures only the nationalistic and related biases of the judges, but should give an indication of how the bias-to-message ratio varies with the extremeness of the message.

Both of these functions are graphed in Figures 1 and 2 for ski jumping and figure skating, respectively. In addition, we graph the probability of the judge being from the same country as the athlete, conditional on the difference between her score and the average of the others. We also graph the probability of a score being influential. A score is regarded as fully influential if both up and down one-increment changes would affect the athlete's score and half influential if either up or down movements would affect the score, but not both. For example, in ski jumping, where the athlete's score is the average of the middle three scores, the second highest score is fully influential, a score that is tied for the highest is half influential, and the highest is not influential.

Figures 1 and 2 reveal that the functions $f()$ and $g()$ are approximately linear even for scores that are up to 1.5 points different from the mean of the other scores in each case. For ski jumping, this is three scoring increments; for figure skating, it is 15. This range of scores contains 99.8 percent of the sample in each sport. Given the aggregate methods used by the sports, scores cease to be influential at much less extreme levels, especially in figure skating. The linearity of $f()$ implies that there is valuable information in these scores that is being truncated, and the linearity of $g()$ implies that the bias-to-message or bias-to-signal ratio of these scores is not higher. Of course, if the aggregation mechanisms increased the weight placed on extreme scores we might also expect the bias to increase, but these results suggest movement in this direction is likely to be optimal.

As mentioned above, $f()$ and $g()$ might appear artificially linear at extreme values if judge's errors or biases were occasionally large in absolute value and if these large errors and the likelihood of making them were correlated for a given performance. Without data on actual performance quality, it is impossible to resolve whether this is a problem in our judging data. What we can do is simulate how misleading the analysis we just performed

in the presence of the above problem.

Figure 3 presents a comparison of a estimate of $f()$ using objective performance quality with one that uses the method of this paper. The data is simulated assuming a panel of 9 judges. Performance quality is distributed $N(0,1)$ and each judge observes this quality plus an i.i.d. noise term that is distributed $N(0,1)$. In addition, with 0.2 probability, 2 of the 9 judges are also affected by an additional noise term, distributed $N(0,4)$. Several aspects of this setup are meant to be fairly extreme in a way that should create problems for the methodology: the fact that only 2 of 9 judges suffer from the large error (to maximize positive skew), that the large error is identical for the two judges, and its magnitude. The dataset we construct includes $50,000$ simulated performances. These parameters produce a distribution of $m_{ipj} - \overline{m}_{ip,-jk}$ with fatter tails than we observe in the data: the simulated kurtosis is 8.5 compared with 6.2 for the ski jumping data and 4.9 for figure skating. And all of the excess kurtosis in the simulated data comes from the large, perfectly correlated errors that affect exactly two of the nine judges so as to create maximum difficulties for our methodology; presumably, the excess kurtosis that exists in the real data is not nearly as problematic.

Figure 3 suggests that whereas the actual $f()$ function would remain roughly linear for score differences of up to 2.9, our method would suggest that it was linear for score differences up to 3.5. In terms of the percent of the sample, our method would lead us to truncate the 2 percent most extreme scores, whereas if we could observe objective performance quality we would decide to truncate the 3 percent most extreme. Notice that while the conclusion about the incremental signal-to-message ratio is different in the 2.9 to 3.5 range, the conclusion about performance quality is roughly the same and only begins to diverge for score differences of greater than 3.8.

Setting truncation points optimally appears to make an economic meaningful difference to the quality of evaluations. If can we compare the mean-squared error of three different aggregation methods using our simulated data: the mean score, the median score, and the mean of all scores with scores more than 2.9 different from the mean of the others truncated at that point. Indexing the MSE of the mean score to 100, the three methods yield MSEs of 100, 70.2, and 71.7. If we exclude the observations with large common errors then the mean minimizes MSE, but truncation at 2.9 still does quite well: 100, 149.7, 100.1. In neither

case does the mean of truncated scores have the lowest MSE, but it appears to be the best method if one is uncertain of the distribution of the data.

3.6 Extreme truncation and vote trading

Two interesting cross-sport correlations emerge from the results above. First, the sports that engage in the most truncation of extreme opinions (figure skating, ski jumping, moguls/aerials, and snowboarding in that order) have the most nationalistically biased judging results. Second, figure skating, a sport with extreme truncation, displays evidence of vote trading, while in ski jumping and moguls, judges appear to compensate for each other's biases.

These correlations do not imply a direction of causality. Figure skating has a long history of suspicion of biased judging, and Campbell and Galbraith (1996) find evidence of a roughly similarly sized bias in 1976, the first Olympics they analyze. The truncation of extreme scores into "votes" is usually justified as an attempt to reduce the extent to which one particular biased judge can influence the results, and so it is possible that the direction of causality is from the behavior of judges to the aggregation method. But it is worth considering whether the truncation may actually be contributing to vote trading, especially since the results of the previous section suggest that a considerable amount of information is being lost through truncation and that signal-to-bias ratios are roughly constant for extreme scores.

There are at least two reasons to worry that it might be. The first is that if the extent to which judges bias their opinions is constrained by reputational concerns, then the truncation of judge's opinions into votes makes the trade-off between the efficacy of biases and their reputational costs very favorable for biasing. To see this, suppose that judges observe $q + n_j$, where q is the true difference between the quality of two performances and n_j is an observational error. Suppose that in one system, judges report $m_j = q + n_j + b_j$, and the evaluation is the average of these reports across judges; whereas in another system, the judges report only the sign of m_j, and the evaluation is determined by majority vote. In the second system, judges report a positive sign unless their observed quality difference is below some threshold: $q + n_j < -b_j$. In the first system, the principal observes a less noisy signal about the b_j used by the judge, $m_j - \widehat{q}$, instead of simply observing whether b_j was greater than of less than $-(q + n_j)$, so updating prior beliefs about judge's biases is

faster, and thus the reputation cost of biasing opinion is higher.

A second reason is that reducing judges' opinions to votes can make reciprocal arrangements easier to sustain. Voting as agreed is a bright line that makes defecting against a reciprocal arrangement easier to detect; with continuous scores, it is more difficult to distinguish observational errors from defection against an agreement to bias a certain amount. The *60 Minutes* tape mentioned in footnote 1, of two judges attempting to signal each other in real time about whether a vote trading agreement still applied, illustrates this difficulty.

4 Implications for judging reform

Following the figure skating judging scandal at the 2002 Olympics, four proposals for judging reform were considered by the International Skating Union (ISU). The proposal that was adopted at a June 2002 ISU meeting was a Canadian proposal to have 14 judges judge each competition, but to only have nine randomly selected judges' rankings count. All 14 scorings would be reported, along with the aggregate ranking, but which nine judges' rankings were used would not be revealed, and which judge issued which scores would also not be revealed. This proposal was adopted in favor of U.S. and Australian proposals that also would randomly select 9 of 14 judgings, but would in addition change the way that they are aggregated. The Australian proposal was to aggregate scores roughly as they are aggregated in ski jumping: ranking skaters based on the mean of the middle five of nine scores. The U.S. proposal was to replace the voting with a ranking based on the median score.

In addition, the ISU agreed to study a longer-term proposal from ISU leadership for replacing the technical merit score with the product of an objective "degree of difficulty" rating and a subjective scoring of execution, similar to the scoring system used in aerials and diving. In addition, the ISU proposed that it, rather than the national federations, select which judges represent particular countries, as the FIS does in skiing.

The results of this paper yield some insights that allow us to comment on these proposals (Table 10). First, the results on judge's reputational concerns in Section 3.4 suggest that allowing a central organization to select judges is likely to yield less biased judges. This is logical, given that the economic interest of the central organization is to maintain viewer interest in the sport (and thus revenue for the organizers), and presumably unbiased judging

is important to doing so.

Second, replacing a completely subjective technical merit score with a partly objective scoring also sounds like a positive change. Aerials, which uses this system, does have the smallest estimated nationalistic bias of the five sports studied.

Third, the Canadian proposal that 14 judges score the competition but only nine judge's scores count, can be considered in three separate stages: 1) expanding the number of judges whose scores may count to 14, 2) using only 9 of the 14 scores, and 3) not revealing which judge issued which scores. The stated reason for expanding the number of potential judges to 14 is to make collusion more difficult to sustain, and economists would generally agree that collusion becomes more difficult as the number of parties increases (e.g., Bain, 1956). Having incurred the costs of using 14 judges, however, the decision to count only nine of the scores is more difficult to rationalize. The results from aggregating 9 of the 14 scores will simply be the results from aggregating all 14 scores plus noise (or 13 out of 14 if an odd number is required). The rationale seems to be that the uncertainty about whether scores will count 1/9 or 0 will lead to less collusion than if all scores counted 1/14, but, if colluding judges care about the expected effect of their collusive arrangement on results, they should be indifferent between these two arrangements.

While it is not clear what is accomplished by not revealing which 9 judgings were used, not revealing which judges issued which scores should make cheating on a collusive arrangement easier, making collusion more difficult to sustain. At the same time, not revealing who issued which score makes it impossible for outsiders to monitor nationalistic biases in judging, and given how we find smaller biases when scrutiny is likely to be higher, a lack of monitoring by outsiders may lead an increase in nationalistic biases. The ISU has addressed this concern by stating that they would keep track of which judge gave which score and review the scores for evidence of judging biases. Whether or not revealing which judge gave which score is a good idea depends on the extent that the ISU can be trusted to pursue this monitoring task vigorously.

Fourth, the Australian proposal would have involved less truncation of extreme scores. The cross-sport evidence suggests that judging biases are smaller and that more likely to cancel each other out in sports that do not truncate extreme scores. Although this relationship is not necessarily causal, the analysis in Section 3.5 suggests that this truncation

involves substantial loss of information, and there are reasons to believe, as outlined in Section 3.6, that truncation encourages vote trading. Perhaps the Australian proposal deserves reconsideration.

5 Implications for organizations

An example of a corporate setting that is roughly analogous to Olympic judging is a promotion committee in a professional services firm, in which senior partners from different offices or practice groups meet to determine which associates to promote. The chair of such a committee faces the problem that the partners who know each associate best are also likely to be biased in their favor. Different committees take different approaches to this problem: excluding partners who are likely to be biased, allowing them to participate but correcting for the likely bias when interpreting their opinions, discouraging biased reporting by the partners by linking their future credibility to their track record for accuracy and unbiasedness, or requiring them to provide evidence, as opposed to just an opinion.

Olympic judging is most similar to a committee that takes the approach of allowing biased partners to participate and offer opinions. The organizers do not adjust for known biases, but, as we have observed in ski jumping, sometimes the other judges do, and the organizers can use career concerns to create incentives for limiting biases.

Unlike in sports judging, most organizations do not insist on simultaneous voting, they instead iterate and attempt to reach a consensus. Formal voting is often viewed as a last resort. The primary effect of iteration is to reduce the role of observational errors, since committee members can condition their final opinions not only on their private information, but also on the opinions of the other committee members. Committee member opinions should differ only if members do not have common priors, if some overweight their private information, or if they differ in their objectives, for example, due to favoritism.

The resulting reduction of observational errors makes biases easier to detect. If an uninformed principal observes two committee members disagreeing, she is quite likely to conclude that one or both of them is biased. The committee members are thus better off agreeing on an immediate opinion, instead of disagreeing, appearing biased, and having the principal average their opinions in some manner anyway. This practice is sometimes called "having the meeting before the meeting," in other words, reaching a consensus privately

instead of airing differences publicly. This process is likely to result in bargaining, with bargaining power depending on who the principal attributes bias to if the parties disagree. A committee member with a professional relationship with the evaluatee may be more likely to have bias attributed to them in the event of a disagreement. This can lead to the effective reclusal of the potentially biased committee member from the decision; she can share her information and opinion with other committee members, but if she is unable to convince them, she is likely to conform to their opinion. On the other hand, she has a strong incentive to seek allies in order to appear less biased, probably in exchange for pushing her allies' candidates, creating incentives for the sort of vote trading we observe in figure skating.

Despite the differences between sports judging and most other team decision making settings, several of the findings in Sections 2 and 3 translate into lessons for organizations.

1. *Valuing fairness.* The existence of compensating biases in ski jumping, and their absence in figure skating suggests that the extent to which judges care about the fairness of results can vary. When judges care about fairness, they compensate for each other's biases, producing evaluations that are less biased than they might otherwise be. To the extent that organizations can cultivate a value of fairness among decision makers, it can lead to nearly unbiased decision making even with biased decision makers.

2. *Using career concerns.* Delegating the selection of judges to interested parties within the organization is likely to produce biased judges, as it appears to in figure skating. A strong but disinterested committee chair, who can adjust the credence paid to members based on their apparent biases and who can create additional incentives for unbiasedness as needed, is likely to improve decision making.

3. *Recognizing the costs of opinion truncation.* The results in Section 3.5 suggest that truncating opinions into votes uses information inefficiently, and the discussion in Section 3.6 suggests that it is likely to encourage bias and vote trading. But truncation may be the only option if a concern for accuracy or reputational concerns fail to restrain opinions, and committee members seek to increase their influence by exaggerating their opinions on every subject. In the absence of a strong chair who can discourage this sort of behavior, there is the possibility for multiple equilibria. In a

good equilibrium, extreme opinions are respected and given higher weight, and committee members police themselves to ensure that they are not extreme too often. They do this to avoid collectively slipping into the bad equilibrium, in which every opinion is extreme, and voting becomes the only way to aggregate opinions. Maintaining the good equilibrium, where information contained in the strength of opinions is not lost, is important, but for the usual reasons, may be impossible in too large a committee.

A Optimal aggregation with normal errors and biases

For simplicity, we will analyze the case with only one judge. With one judge, the principal's aggregate problem is one of deciding how much to update her prior belief about performance quality q in response to an extreme or less-extreme opinion. Assume that observational errors and biases are normal and i.i.d.: $E(q|s) = q + e$, $q \sim N(0,1)$, $e \sim N(0, V_e)$ and $b \sim N(0, V_b)$.

The principal's problem is to

$$\max_{y()} E[-(y\{m[s,b,y()]\} - q)^2],$$

with the agents opinion given by her first order condition as in Section 2:

$$m = q + e + b \cdot y'(m).$$

We will assume that this expression uniquely defines m for a given q, e, and b. This requires that y' does not increase rapidly for extreme scores, which turns out to be the case. The principal's objective function can be written:

$$\begin{aligned} & \min_{y()} \iiint [y(m) - q]^2 \cdot f(m,q,b) \cdot dq \cdot db \cdot dm \\ f(m,q,b) &= \phi(q) \cdot \phi(\frac{b}{V_b^{1/2}}) \cdot \phi[\frac{m - q - b \cdot y'(m)}{V_e^{1/2}}], \end{aligned}$$

where $\phi()$ is the standard normal p.d.f. This yields a first order condition for each $y(m)$. We can instead write the FOC for an increment to $y'(m)$:

$$\begin{aligned} -2 \int_m^\infty \int_b \int_q [y(\widetilde{m}) - q] \cdot f(q,b,\widetilde{m}) &= \int_b \int_q [y(m) - q]^2 \cdot \frac{df(m,q,b)}{dy'(m)} \\ &= \int_b \int_q [y(m) - q]^2 \cdot \frac{b \cdot e}{V_n} \cdot f(m,q,b), \end{aligned}$$

which must hold for all m. Differentiating this expression by m yields:

$$\begin{aligned} \int_b \int_q [y(m) - q] \cdot f(q,b,m) &= \int_b \int_q [y(m) - q] \cdot \frac{2 \cdot y'(m) \cdot b \cdot e + [y(m) - q] \cdot [b - b^2 \cdot y''(m)] \cdot (1 - e^2 \cdot V_e^{-1})}{2V_e} \cdot f(\\ y(m) &= E(q|m) + \frac{E\{[y(m) - q] \cdot 2 \cdot y'(m) \cdot b \cdot e | m\}}{2V_e} \\ &\quad + \frac{E\{[y(m) - q]^2 \cdot b \cdot (1 - e^2 \cdot V_e^{-1}) | m\}}{2V_e} \\ &\quad - \frac{E\{[y(m) - q]^2 \cdot b^2 \cdot y''(m) \cdot (1 - e^2 \cdot V_e^{-1}) | m\}}{2V_e} \end{aligned}$$

Solving this expression for $y(m)$ in closed form difficult, but we can use it to describe $y(m)$. The first term in the expression above yields a linear scheme, $y(m) = E(q|m)$. We are interested in whether the slope $y'(m)$ increases or decreases with the absolute value of m, and thus in whether $y''(m)$ has the same sign as m (overweighting extreme opinions) or the opposite sign (underweighting extreme opinions) as opinions become extreme. In other words, we are interested in the sign of $y'''(m) > 0$ as m becomes large.

The second term has an unambiguously positive third derivative. The third term has a positive third derivative until the expected value of e given m becomes large relative to its variance – or equivalently, until m becomes large relative to its variance, after which it has a negative third derivative. The fifth derivative of the third term is unambiguously negative, so as m becomes extreme, $y'(m)$ will change to the opposite sign of m. The fourth term acts to dampen any curvature in $y(m)$ for values close to zero and reinforce it for values further away, since the third derivative of the term will be have the opposite sign as y'' for values close to zero and the same sign as m becomes large relative to its variance.

Taking these terms together suggests that $y(m)$ will have a positive third derivative until the absolute value of m is some critical value, after which it will have a negative third derivative. Simulations for values of V_b and V_e between 0.25 and 4 suggest that this causes the sign of $y'(m)$ to change as m is about two standard deviations from its mean. Until this point, $y'(m)$ is very close to linear.

Since the optimal mechanism seems to be approximated by linearity up to a critical value, we can approximate it by analyzing this more tractable class of mechanisms. In particular, we now assume that $y(m)$ is $\alpha \cdot m$, but its absolute value is constrained to be less than c.

In response to this mechanism, the judge will report $m = q + e + b \cdot \alpha$, or $\pm c \cdot a^{-1}$ whichever is closer to zero. The principal chooses α and c to maximize:

$$\max_{\alpha,c} -\int_q \int_b \{\int_{m=-c\alpha^{-1}}^{c\alpha^{-1}} (\alpha m - q)^2 \cdot f(m,q,b) - 2\int_q \int_b \int_{m=c\alpha^{-1}}^{\infty} (c-q)^2 \cdot f(m,q,b).$$

The expectation $E(q|m) = m(1+\alpha^2 \cdot V_b + V_e)^{-1} = \gamma \cdot m = m \cdot V_m^{-1}$. The first order condition for increasing c is:

$$E(q|m > c\alpha^{-1}) - c = E(q|E(q|m) > c\frac{\gamma}{\alpha}) - c.$$

This implies $c = \infty$ unless $\alpha > \gamma$. To get a first order condition for α, we rewrite the first part

of the maximization expression can be written $E(m^2||m| < c\alpha^{-1}) \cdot (\alpha-\gamma)^2 + Var(q|m, |m| < c\alpha^{-1})$ and then differentiate:

$$\alpha = \gamma - \frac{1}{2 \cdot E(m^2||m| < c\alpha^{-1})} \cdot [\frac{dE(m^2||m| < c\alpha^{-1})}{d\alpha} \cdot (\alpha - \gamma)^2 + \frac{dVar(q|m, |m| < c\alpha^{-1})}{d\alpha}].$$

The variance $Var(q|m) = (1-\gamma)^2 \cdot Var(m) = (1-\gamma)^2\gamma^{-1}$ for all m. Its derivative with respect to α is $2\alpha V_b(1-\gamma^2)$. Since the other term is zero when $\alpha = \gamma$ and increases with α quadratically, this implies that α will be less than γ. This in turn implies that the optimal c is infinite, so the condition above reduces to

$$\begin{aligned} \alpha &= \gamma - \alpha V_b(1-\gamma^2)\gamma^{-1} \\ &= \gamma \cdot [1 + V_b \cdot (V_m - \gamma)]^{-1}. \end{aligned}$$

The incentive scheme is linear, its slope is less than without pre-commitment, and the slope decreases with V_b.

Taken together, these analyses imply that any deviation from linearity should involve some truncation of extreme scores, but that if the principal is limited to a linear scheme with truncation at a critical value, then truncation is no longer optimal. This suggests that the commitment value in truncating extreme scores is probably minimal, so long as signal-to-bias and signal-to-noise ratios do not decline for extreme scores.

References

Aghion, Phillipe and Jean Tirole, "Formal and Real Authority in Organizations," *Journal of Political Economy*, 1997, *105*, 1–29.

Athey, Susan and John Roberts, "Organizational Design: Decision Rights and Incentive Contracts," *American Economic Association Papers and Proceedings*, 2001, *91*, 200–205.

Bain, Joseph, *Barriers to New Competition*, Cambridge, MA: Harvard University Press, 1956.

Bassett, Gilbert W. and Joseph Persky, "Rating Skating," *Journal of the American Statistical Association*, 1994, *89*, 1075–1079.

Bronars, Stephan and Gerald Oettinger, "Performance, Participation and Risk-Taking in Tournaments: Evidence from Professional Golf," 2001. University of Texas Mimeo.

Campbell, Bryan and John W. Galbraith, "Non-parametric tests of the unbiasedness of Olympic figure-skating judgments," *The Statistician*, 1996, *45*, 521–526.

Duggan, Mark and Steven D. Levitt, "Winning Isn't Everything: Corruption in Sumo Wrestling," 2000. NBER Working Paper 7798.

Garicano, Luis, Ignacio Palacios, and Canice Prendergast, "Favoritism Under Social Pressure," 2001. University of Chicago Mimeo.

Goff, Brian L., Robert E. McCormick, and Robert D. Tollison, "Racial Integration as an Innovation: Empirical Evidnece from Sports Leagues," *American Economic Review*, 2002, *92*, 16–26.

Meyer, Margaret, Paul Milgrom, and John Roberts, "Organizational Prospects, Influence Costs, and Organizational Change," *Journal of Economics and Management Strategy*, 1992, *1*, 9–35.

Milgrom, Paul and John Roberts, "Relying on the Information of Interested Parties," *RAND Journal of Economics*, 1986, *17*, 18–32.

Prendergast, Canice, "A Theory of "Yes Men"," *American Economic Review*, 1993, *83*, 757–770.

___ and Robert Topel, "Favoritism in Organizations," *Journal of Political Economy*, 1996, *104*, 958–978.

Romer, David, "It's Fourth Down and What Does the Bellman Equation Say? A Dynamic-Programming Analysis of Football Strategy," 2002. NBER Working Paper No. 9024.

Wolfers, Justin, "Point Shaving: Corruption in NCAA Basketball," 2002. Stanford Univeristy Mimeo.

Zitzewitz, Eric, "Measuring herding and exaggeration by equity analysts," 2001. Stanford Univeristy Mimeo.

___, "Regulation Fair Disclosure and the Private Information of Analysts," 2002. Stanford University Mimeo.

Table 1. Dataset dimensionality

	Figure skating	Ski jumping	Sport Moguls	Aerials	Snowboarding
Events	61	25	16	16	15
Olympics	4	3	2	2	2
Pre-Olympics, Olympic-level	41	17	10	14	5
Post-Olympics, Olympic-level	0	5	4	0	6
Junior level	16	0	0	0	2
Rounds/jumps	181	62	28	42	31
Compulsory (skating)/Qualifying (skiing)	59	12	11	10	0
Short/Long programs (skating)/Finals (skiing)	122	50	17	32	31
Performances	2,976	2,920	1,016	1,067	643
Scores	25,068	14,600	7,112	7,469	3,147
Unique athlete countries	54	28	21	18	28
Unique athletes	584	243	125	108	249
Performances per athlete	5.1	12.0	8.1	9.9	2.6
Scores per performance	8.4	5.0	7.0	7.0	4.9
Unique judge countries	41	15	15	11	9
Unique judges	314	75	31	15	12
Events per judge	2.0	1.7	3.6	7.5	6.1

Table 2. Summary statistics for scores

Sport	Score type	Number of scores per performance	Range	Minimum increment	Mean	Standard deviation	
						Overall	Within performance
Figure skating	Technical Merit (TM)	5, 7, or 9	0 to 6	0.1	4.48	0.69	0.20
	Artistic impression (AI)	5, 7, or 9	0 to 6	0.1	4.72	0.61	0.19
	Total (TM + AI)	5, 7, or 9	0 to 12	0.1	9.20	1.28	0.35
	Ordinal placement	5, 7, or 9	1 to 32	1.0	11.11	7.11	1.59
Ski jumping	Style points	5	0 to 20	0.5	17.57	1.09	0.33
Moguls	Turns	5	0 to 5	0.1	4.00	0.78	0.18
	Air	2	0 to 7.5	0.01	4.84	1.22	0.19
Aerials	Air & Form	5	0 to 7	0.1	5.61	1.43	0.25
	Landing	2	0 to 3	0.1	2.01	0.98	0.06
Snowboard halfpipe	Standard Air	1	0 to 10	0.1	5.51	1.97	NA
	Rotations	1	0 to 10	0.1	5.39	2.45	NA
	Amplitude	1	0 to 10	0.1	5.53	1.97	NA
	Overall impression	2	0 to 10	0.1	5.34	2.09	0.18
	All scores	5	0 to 10	0.1	5.43	2.12	1.08

Notes:

1. In figure skating, each of 5, 7 or 9 judges issues scores for both technical merit and artistic impression. In moguls, aerials, and snowboarding, judges are assigned different aspects of the performance to judge, and thus their scores are not necessarily comparable across aspects.
2. The within performance standard deviation of all scores for snowboarding allows for fixed effects for score type.

Table 3. Nationalistic bias by sport

Sport	Performances	Scores	Estimated bias in points		In standard deviations	
			Coeff.	S.E.	Overall	Within perf.
Figure skating						
Technical merit (TM)	2,976	25,068	0.077	0.006	0.111	0.394
Artistic impression (AI)	2,976	25,068	0.089	0.005	0.147	0.482
Technical merit + Artistic Impression (TM + AI)	2,976	25,068	0.166	0.010	0.130	0.474
Ordinal placement	2,976	25,068	-0.704	0.045	-0.099	-0.443
Ski jumping	2,920	14,600	0.145	0.011	0.132	0.443
Mogul skiing						
Turns	1,016	5,080	0.084	0.010	0.108	0.479
Air	1,016	2,032	0.053	0.032	0.043	0.270
Aerials						
Air & Form	1,067	5,335	0.014	0.012	0.010	0.057
Landing	1,067	2,134	0.010	0.009	0.010	0.158
Snowboard Halfpipe	643	3,147	0.099	0.091	0.047	0.092

Notes:

1. Nationalistic bias is estimated by regression scores on a dummy variable for the judge and athlete being from the same country, plus performance and judge country fixed effects (the same specification as in Table 4, Line 1).

Table 4. Alternative specifications and identification approaches

Dependent variable: style/turns point score (ski jumping and moguls) or TM+AI score total (figure skating)

Panel A. Ski jumping

					Same country		Different country	
Line	Performance Effects	Skier FEs	Jump characteristics	Obs.	Coeff.	S.E.	Coeff.	S.E.
(1)	Fixed	No	No	14,600	0.145	0.011		
(2)	Random	No	No	14,600	0.148	0.011		
(3a)	Random	Yes	No	14,600	0.146	0.011		
(3b)	Random	Yes	Yes	14,600	0.144	0.011		
(4a)	Random	Yes	No	14,600	0.104	0.036	-0.040	0.015
(4b)	Random	Yes	Yes	14,600	0.123	0.038	-0.021	0.017

Panel B. Figure skating (TM + AI)

					Same country		Different country	
Line	Performance Effects	Skater FEs	Objective characteristics	Obs.	Coeff.	S.E.	Coeff.	S.E.
(1)	Fixed	No	N/A	25,068	0.166	0.010		
(2)	Random	No	N/A	25,068	0.169	0.010		
(3)	Random	Yes	N/A	25,068	0.167	0.010		
(4)	Random	Yes	N/A	25,068	0.258	0.026	0.092	0.025

Panel C. Figure skating (Ordinal placement)

					Same country		Different country	
Line	Performance Effects	Skater FEs	Objective characteristics	Obs.	Coeff.	S.E.	Coeff.	S.E.
(1)	Fixed	No	N/A	25,068	-0.704	0.045		
(2)	Random	No	N/A	25,068	-0.712	0.045		
(3)	Random	Yes	N/A	25,068	-0.702	0.045		
(4)	Random	Yes	N/A	25,068	-2.400	0.218	-1.696	0.215

Panel D. Mogul skiing (Turns)

					Same country		Different country	
Line	Performance Effects	Skater FEs	Run characteristics	Obs.	Coeff.	S.E.	Coeff.	S.E.
(1)	Fixed	No	No	5,080	0.084	0.010		
(2)	Random	No	No	5,080	0.085	0.010		
(3a)	Random	Yes	No	5,080	0.084	0.010		
(3b)	Random	Yes	Yes	5,080	0.083	0.010		
(4a)	Random	Yes	No	5,080	-0.044	0.058	-0.128	0.077
(4b)	Random	Yes	Yes	5,080	-0.031	0.043	-0.115	0.062

Notes:

1. All regressions include judge country fixed effects.
2. To capture differences in conditions, judging standards, etc., the regressions that use performance random effects include fixed effects for meet*round combinations.
3. Regressions that control for jump and run characteristics include distance jumped and takeoff speed (for ski jumping) and time (for moguls), and interactions of these variables with meet fixed effects.

Table 5. Nationalistic biases in subsamples of the data

Panel A. Ski jumping

	Obs.	Nationalistic bias Coeff.	S.E.	p-value of bias difference with next category
All	11,670	0.155	0.013	
Olympics	1,945	0.258	0.033	0.000
Non-Olympics (World Cup events)	9,725	0.137	0.014	
Final round	7,010	0.147	0.015	0.474
Qualifying round	3,215	0.149	0.025	
Final round, top 10 finisher	1,710	0.212	0.028	0.003
Final round, not top 10 finisher	5,300	0.121	0.018	
Qualifying round, not-pre qualified	2,855	0.153	0.027	0.240
Qualifying round, pre-qualified	360	0.107	0.059	
Team competition	1,445	0.218	0.043	0.057
Individual competition	10,225	0.147	0.013	
Individual competition, K90 hill	1,725	0.232	0.037	0.006
Individual competition, K120 hill	8,500	0.133	0.014	

Panel B. Figure skating (TM+AI)

	Obs.	Nationalistic bias Coeff.	S.E.	p-value of bias difference with next category
All	25,068	0.166	0.010	
Olympics	2,134	0.128	0.028	0.114
Non-Olympics, senior	16,643	0.165	0.012	0.262
Junior	6,291	0.180	0.021	
Long program	9,145	0.140	0.013	0.065
Short program	10,050	0.172	0.017	
Long or short program, top 10 finisher	9,951	0.141	0.012	0.038
Long or short program, not top 10 finisher	9,244	0.183	0.020	
Ice dancing	8,166	0.198	0.016	0.007
Men's, Women's, or Pairs	9,388	0.148	0.013	
Women's	7,136	0.174	0.021	0.109
Men's	6,597	0.137	0.022	0.256
Pairs	2,791	0.116	0.023	
Ice dancing, short or long	4,519	0.179	0.017	0.091
Ice dancing, complusories	5,873	0.224	0.029	
Technical Merit (TM)	25,068	0.077	0.006	0.052
Artistic impression (AI)	25,068	0.089	0.005	

Notes:

1. This table replicates the regression in Table 4, Line 1 for subsamples of the data.
2. To avoid a sample selection bias, which athletes are "top 10" or "not top 10" is determined by replacing the score in the current observation and then reranking the competitors in that contest.

Table 6. Nationalistic bias and leniency by judge country

This table reports average nationalistic biases and softness by country of judge affiliation. Leniency is measured using judge country fixed effects in a regression using the same specification as in Line 1 of Table 4 (i.e. with performance fixed effects); nationalistic bias is the interaction of the judge country fixed effects with a dummy variable for the athlete being from the same country as the judge. Only the 25 countries with the most same-country athlete observations are shown for figure skating; all 15 countries are shown for ski jumping.

Panel A. Ski jumping

		Nationalistic bias		Leniency		Observations	
Country	Abrev.	Coeff.	S.E.	Coeff.	S.E.	All scores	Same-country athlete
Korea	KOR	0.386	0.203	0.175	0.049	80	4
Slovakia	SVK	0.297	0.121	0.071	0.025	617	11
France	FRA	0.292	0.111	0.017	0.026	583	13
Czech Republ	CZE	0.255	0.082	0.000	0.021	467	25
Slovenia	SLO	0.249	0.041	0.002	0.022	1,279	104
Sweden	SWE	0.246	0.230	-0.101	0.025	610	3
Germany	GER	0.180	0.027	0.079	0.020	2,046	245
Austria	AUT	0.153	0.027	0.103	0.020	2,246	253
Poland	POL	0.145	0.057	0.091	0.026	555	53
Italy	ITA	0.144	0.100	0.119	0.024	762	16
Norway	NOR	0.109	0.033	-0.093	0.021	1,839	158
Finland	FIN	0.081	0.034	-0.035	0.023	1,256	153
Japan	JPN	0.041	0.035	0.060	0.024	900	159
Switzerland	SUI	0.024	0.077	0.008	0.025	542	28
USA	USA	0.008	0.078	0.014	0.023	818	27

Panel B. Figure skating

		Nationalistic bias		Leniency		Observations	
Country	Abrev.	Coeff.	S.E.	Coeff.	S.E.	All scores	Same-country athlete
Azerbaijan	AZE	0.316	0.075	0.001	0.051	699	28
Hungary	HUN	0.310	0.068	-0.064	0.051	826	34
Slovenia	SLO	0.306	0.113	0.057	0.052	480	12
Romania	ROM	0.300	0.109	0.024	0.052	542	13
Korea	KOR	0.290	0.124	0.018	0.055	258	10
Slovakia	SVK	0.248	0.079	-0.015	0.051	739	25
Uzbekistan	UZB	0.231	0.095	0.016	0.055	260	18
Poland	POL	0.225	0.049	-0.033	0.050	929	66
Canada	CAN	0.210	0.032	-0.108	0.050	1,610	158
Italy	ITA	0.205	0.046	-0.083	0.050	1,167	76
Czech Republ	CZE	0.174	0.058	-0.029	0.051	889	47
USA	USA	0.172	0.033	-0.087	0.050	1,387	159
Belgium	BEL	0.156	0.124	-0.064	0.053	392	10
Germany	GER	0.155	0.045	-0.016	0.050	1,423	80
Finland	FIN	0.144	0.063	-0.055	0.051	790	39
China	CHN	0.132	0.057	0.082	0.053	454	52
Australia	AUS	0.132	0.051	-0.076	0.051	1,077	60
Japan	JPN	0.129	0.044	0.037	0.050	1,238	83
Russia	RUS	0.117	0.030	0.063	0.050	1,603	190
Bulgaria	BUL	0.115	0.079	-0.028	0.051	698	25
France	FRA	0.114	0.038	-0.037	0.050	1,156	114
Switzerland	SUI	0.114	0.060	-0.017	0.050	1,096	43
Ukraine	UKR	0.110	0.040	0.000	0.050	1,331	101
Estonia	EST	0.063	0.074	0.002	0.052	575	29
Great Britain	GBR	0.041	0.074	-0.238	0.051	702	29

Table 7. Variation of compensating bias with the nationalistic bias of other judges
Dependent variable: style point score (ski jumping) or TM+AI score total (figure skating)

	Ski jumping		Figure skating	
	Coeff.	S.E.	Coeff.	S.E.
Athlete country same as judge	0.101	0.046	0.280	0.030
Athlete country represented on panel	0.086	0.051	0.256	0.030
(Athlete country represented on panel)*(Athlete country judge bias)	-1.259	0.213	-0.998	0.103
Athlete finishes within two places of athlete from judge country	0.004	0.015	0.005	0.015
Observations	10,100		23,890	

Notes:

1. Each column is a regression. Regressions include performance random effects and fixed effects for athletes and meet*event*round combinations (I.e., they use same specification as Line 4 in Table 4).

Table 8. Bias matrix for figure skating

Average d-score (i.e., difference between score and the average of the other scores given that performance) for judge-athlete country combination less average d-score for judge country. Results are for the sum of TM and AI scores.

Panel A. Bias matrix

	Judge country									
Athlete country	CAN	USA	GER	ITA	JPN	CHN	RUS	UKR	FRA	POL
CAN	0.150	0.036	0.034	0.085	-0.016	-0.039	-0.018	-0.047	-0.044	-0.012
USA	0.032	0.125	-0.026	-0.048	-0.039	0.009	-0.061	0.014	0.012	-0.065
GER	-0.079	-0.004	0.154	0.033	0.012	-0.044	-0.029	-0.018	-0.165	-0.008
ITA	-0.051	-0.030	0.004	0.130	-0.077	0.087	0.038	0.080	-0.075	-0.054
JPN	0.074	-0.023	0.003	-0.024	0.112	-0.031	-0.086	-0.071	-0.005	0.043
CHN	-0.016	0.003	0.022	-0.001	-0.033	0.134	-0.099	-0.028	0.017	0.042
RUS	-0.023	-0.063	-0.002	-0.040	-0.014	-0.035	0.104	-0.030	0.008	-0.061
UKR	-0.126	-0.048	-0.052	-0.008	-0.031	-0.111	0.036	0.113	0.027	-0.058
FRA	-0.008	-0.068	-0.076	0.005	-0.008	-0.024	-0.024	-0.059	0.091	0.040
POL	-0.238	-0.227	-0.066	-0.139	0.152	0.008	-0.029	0.030	0.098	0.176

Panel B. Summary by bloc

	Judges from		
Average bias	Bloc A	Bloc B	Neither
For own athletes	0.140	0.121	0.123
For athletes in Bloc A (CAN, USA, GER, ITA)	-0.001	-0.028	-0.013
For athletes in Bloc B (RUS, UKR, FRA, POL)	-0.074	-0.002	-0.008
For athletes in neither Bloc (JPN, CHN)	0.005	-0.023	-0.032

Table 9. Characteristics of ski jumping judges chosen to judge in the Olympics

This table compares the pre-Olympic judging history of judges that were chosen to judge in the Olympics with those that were not. Judges are compared on leniency (their average d-score, the difference between their score and the average score given a performance), nationalism (the difference between the average d-score for compatriots and the average for all athletes), and deviation from other judges (the average absolute d-score). All standard errors and p-values reported are heteroskedasticity robust. Marginal probit coefficients are reported; asterisks indicate significance at the 5 percent level.

Panel A. Softness and nationalistic bias of judges chosen and not chosen to judge in the Olympics

	(1) Chosen for Olympics		(2) Not chosen		(3) Not chosen from country with judge		P-values for comparing means	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	(1) vs. (2)	(1) vs. (3)
Ski jumping	N = 3		N = 47		N = 32			
Leniency (average d-score)	-0.046	0.127	0.002	0.096	0.004	0.107	0.449	0.405
Nationalism (average d-score for own-country athletes less average d-score)	0.010	0.011	0.128	0.121	0.100	0.082	0.000	0.000
Deviation (average absolute d-score)	0.235	0.050	0.244	0.050	0.253	0.051	0.730	0.454
Figure skating	N = 30		N = 165		N = 130			
Leniency (average d-score, TM and AI combined)	0.040	0.114	-0.019	0.114	-0.019	0.112	0.006	0.008
Nationalism (average d-score for own-country athletes less average d-score)	0.207	0.148	0.131	0.210	0.119	0.192	0.006	0.016
Deviation (average absolute d-score)	0.247	0.059	0.256	0.063	0.254	0.065	0.401	0.540

Panel B. Probit regressions predicting selection to judge in the Olympics

	Obs.	Leniency	Nationalism	Deviation
Ski jumping				
Including all countries' judges	50	0.044	-0.171*	-0.128
		(0.127)	(0.129)	(0.135)
Including six countries with judges at Olympics (AUT, GER, JPN, NOR, SLO, USA)	35	0.052	-0.219*	-0.211
		(0.157)	(0.242)	(0.242)
Figure skating				
Including all countries' judges	195	0.482	0.128	-0.508
		(0.307)	(0.117)	(0.423)
Including 19 countries with judges at Olympics	160	0.503	0.268	-0.558
		(0.344)	(0.160)	(0.479)
Including country fixed effects	160	0.645	0.296	-0.842
		(0.409)	(0.197)	(0.539)

Table 10. Implications of the results of the paper for current proposals for reforming figure skating judging

Proposal and aspect	Adopted in June 2002?	Implications of paper's findings for desirability	
		Sign	Rationale
ISU proposal			
ISU selects individual judges, rather than national federations	Yes	+	Career concerns results suggest that FIS chooses less biased judges in ski jumping, while figure skating national federations choose more biased judges.
Technical merit scoring replaced with objective degree of difficulty measure, multiplied by score for execution	Tabled for further study	+	Less nationalistic bias in aerials, which uses similar system. Less bias for technical merit than for artistic impression, which is more objective.
Canadian proposal	Yes		
Have 14 judges instead of 9		+	Increasing number makes vote trading more difficult to implement
Randomly select 9 out of 14 scores to count		-	Adds noise to results, relative to using 13 or 14. Should not deter collusion if judges are risk neutral.
Reveal all 14 scores, but not which were used		?	
Do not reveal which judge gave which scores		+/-	Should reduce collusive agreements harder by making defection from them easier, but we would then need to trust the ISU to monitor judges for bias.
Australian proposal	No		
Rank skaters using mean of middle 5 scores instead of voting		+	Truncation of extreme scores leads to loss of information and may help facilitate vote trading. Less nationalistic bias in sports that truncate less.
U.S. proposal	No		
Rank skaters using median score instead of voting		Weakly +	Involves slightly (but only slightly) less truncation of extreme scores than current system.

Figure 1. Signal and bias content of extreme judge opinions -- ski jumping

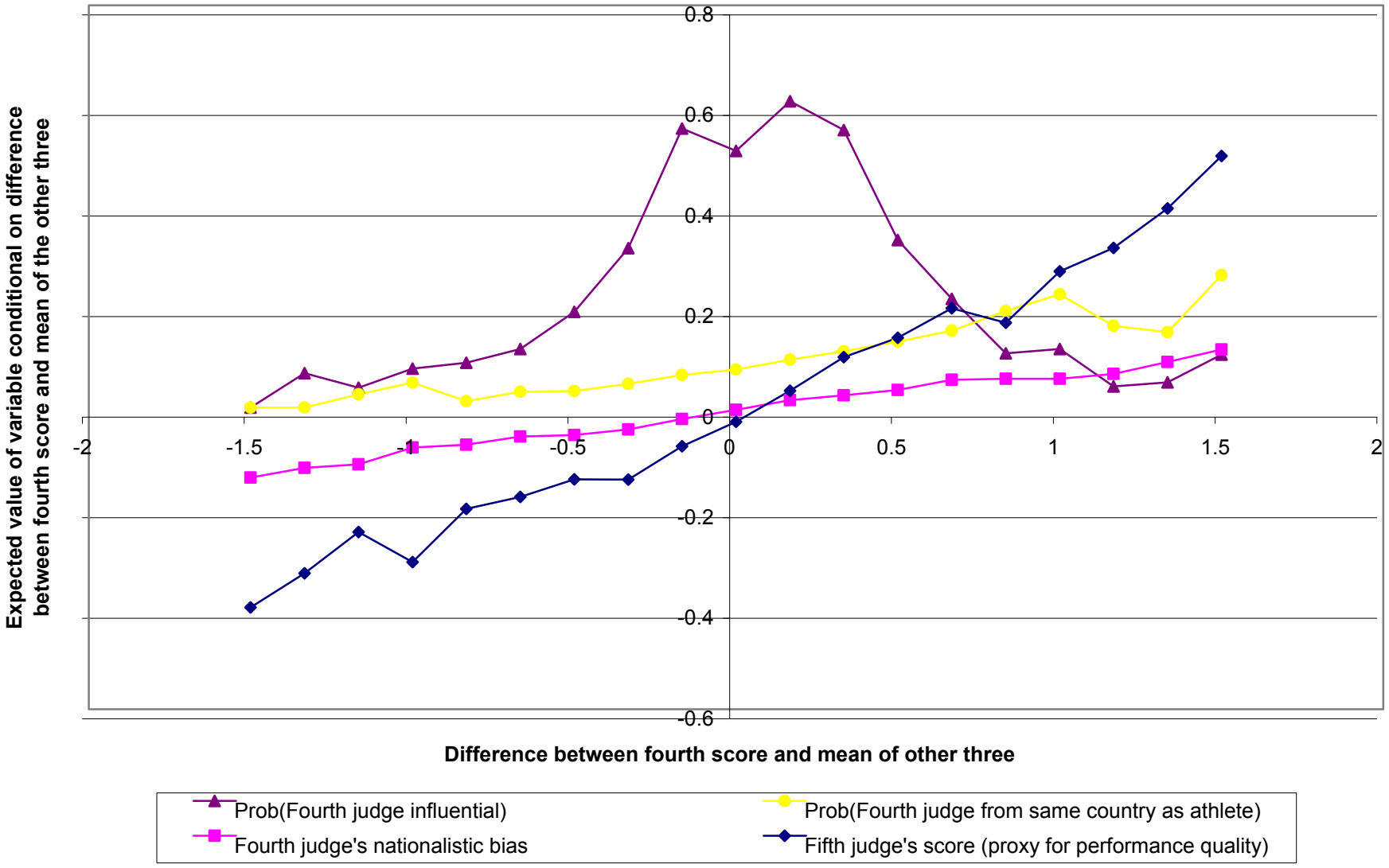
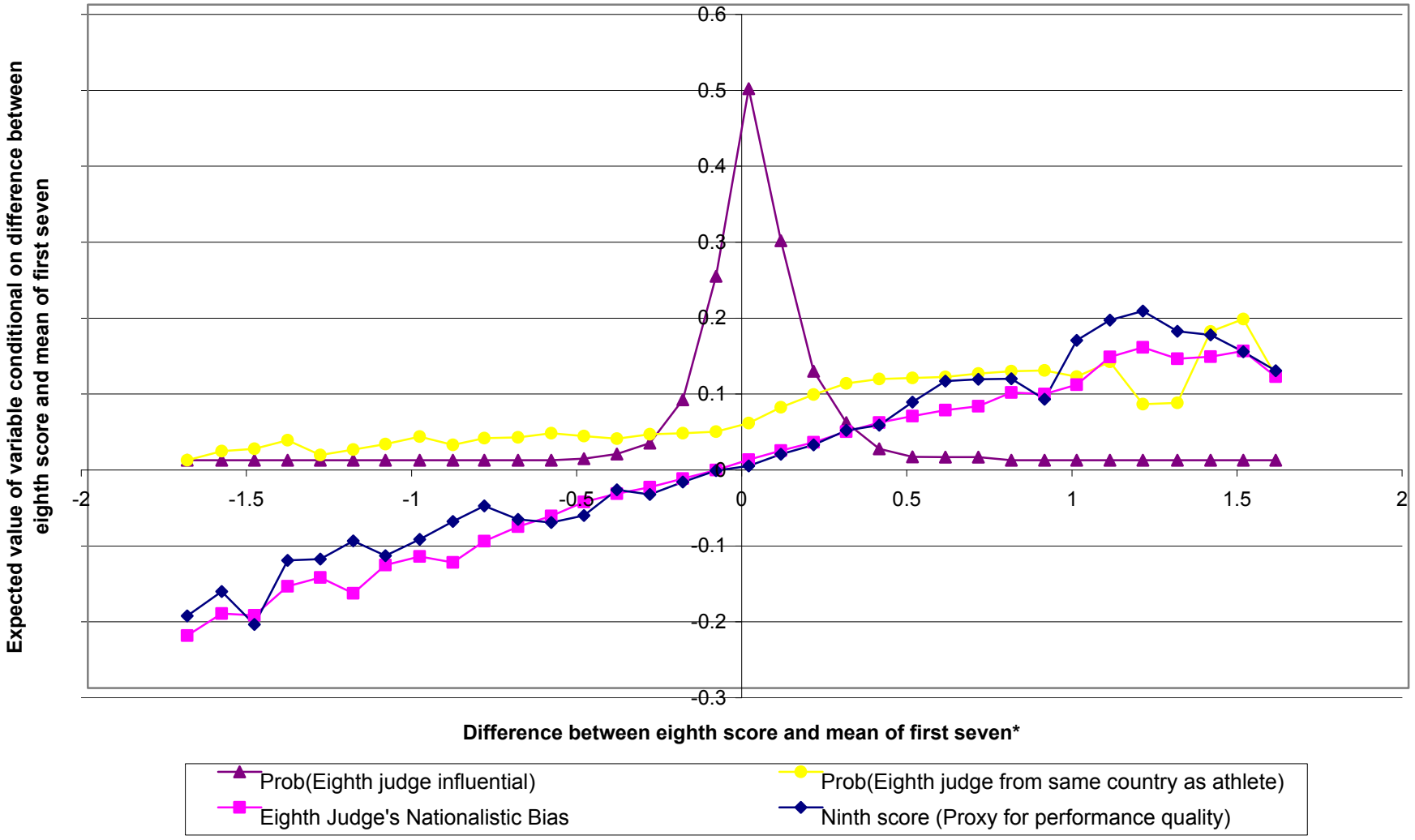


Figure 2. Signal and bias content of extreme judge opinions -- figure skating



* For simplicity, this analysis includes only events judged by nine judges, which are about 80 percent of the sample.

Figure 3. Comparison of approximate and ideal method for simulated data