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

Eric Zitzewitz
Stanford Graduate School of Business
Stanford, CA 94305
ericz@stanford.edu

This paper exploits nationalistic biases in Olympic winter sports 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 subjectivity 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. Extreme truncation not only discards information, but may also make the vote trading in figure skating easier to implement. 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 routinely 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

The author thanks Severin Borenstein, Gary Charness, Judy Chevalier, Kenneth Fuller, Robert Gibbons, Benjamin Hermalin, Phil Leslie, Steve Levitt, John Morgan, Paul Oyer, Canice Prendergast, John Roberts, Scott Stern, Alan Sorenson, Brian Viard, Justin Wolfers, two anonymous referees, three anonymous former athletes and judges, and seminar and conference participants at Berkeley, George Washington, HBS, and the NBER for helpful suggestions and comments.

Journal of Economics & Management Strategy, Volume 15, Number 1, Spring 2006, 67–99

or project best are often those most likely to be biased. In deciding how much to count on 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 one particular team decision-making setting. In order to do meaningful empirical work about biases that one might expect to be fairly subtle, one needs 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 I examine an analogous setting in sports: the judging of winter Olympic sports.

Olympic judges represent a particular country and, in the sports I study, they display biases in favor of athletes from the same country. These nationalistic biases can be reasonably large. Using individual judges' scoring data from the 2002 Olympics and other major international competitions, I find that figure skating and ski jumping judges each score their compatriots about 0.13 standard deviations higher than other judges.[1] In figure skating, where placement is determined entirely by subjective judging, this bias translates to an average placement 0.7 positions higher. These biases appear even larger when compared to the standard deviation of scores awarded to a particular performance; they are about 45% of the within-performance standard deviation of scores in both sports. Biases this large can be statistically identified in a sample size of about 20; thus many armchair empiricists are already aware of a nationalistic bias in figure skating, at least.

In most settings, attempts to study favoritism empirically would be frustrated by the difficulty of observing where one should expect favoritism (e.g., who is "friends" with whom). In this study, judges are biased nationalistically, and thus in at least one way that a researcher

[1] Data were also collected for the other judged Olympic winter sports (mogul skiing, aerials, and snowboarding halfpipe). These sports also display nationalistic judging biases, albeit smaller ones. The sample sizes for these sports unfortunately do not allow for cross-sectional analyses, however, and so I am omitting them from this version of the paper for space reasons (see Zitzewitz, 2002, for these results).

can observe. This allows me to study how the degree of favoritism varies with strategic considerations and how organizational design can be effective or ineffective in dealing with favoritism.

There are three basic findings. First, judges vary the magnitude of their nationalistic biases in response to their career concerns. The opportunity to judge in the Olympics is by far the greatest reward available to a judge. In ski jumping, judges are chosen by a central body, the judges subcommittee of the *Federation International du Ski* (FIS). Consistent with its primary interest of preserving the integrity of and interest in the sport, the committee chooses judges who have been less biased than average in judging pre-Olympic events, controlling for other measures of judging performance. Consistent with the incentives this creates, ski jumping judges are less nationalistically biased at pre-Olympic events than they are at the Olympics. In figure skating, in contrast, national federations choose the judges to represent them at the Olympics. They choose the most nationalistically biased, as one would expect given their incentives.[2] Figure skating judges actually display greater nationalistic biases in pre-Olympic competitions—perhaps because of the absence of media scrutiny, or perhaps because their career concerns drive them to be *more* biased than they would be otherwise.[3]

Second, judges' biases interact strategically. In ski jumping, judges compensate for the biases of a particularly nationalistic colleague, leading results to be more fair than they would otherwise be. A ski jumper is actually hurt slightly by having a compatriot on the judging panel. In figure skating, in contrast, judges reinforce each other's biases and appear to engage in bloc judging or vote trading. A skater whose country is not represented on the judging panel is at a serious disadvantage. The data suggest that countries are divided into two blocs, with the US, Canada, Germany, and Italy on one side and Russia, the Ukraine, France, and Poland on the other. The judging blocs are thus what they appeared to be in the controversial 2002 Olympics pairs competition, even for events 1–2 years before the Olympics. That, together with the fact that the judging blocs seem more reminiscent of Cold War era geopolitical relationships than of contemporary relationships or nationalistic sentiment, suggests that vote trading relationships are fairly long-lived.

[2] Perhaps the best recent example is that of Ukrainian judge Yuri Balkov who, prior to the 1998 Olympic Ice Dancing competition, was taped by another judge announcing the order in which he would rank contestants. Balkov was suspended by the ISU for three years, but was subsequently chosen by the Ukraine to judge in the 2002 Olympics.

[3] In her initial comments following the 2002 Pairs Olympic competition, French judge Marie-Reine Le Gougne indicated that she was pressured by her national federation to vote for the Russian couple, suggesting that the national federation was pushing a judge to be more biased than she wanted to be. She did subsequentially change her story, however.

Third, the current methods for aggregating judges' opinions appear to place too little weight on extreme opinions. In ski jumping, five judges score each jump on style, and the skier's score is the sum of the middle three scores. This method involves some truncation of extreme scores, but the likelihood that a judge's score will be influential remains high. In figure skating, the relative ranking of a pair of skaters is determined by which skater is ranked higher on a majority of score cards.[4] This represents the most aggressive truncation of opinions about the relative quality of two skaters possible. Truncation of extreme opinions is usually justified as a means of reducing the influence of the most biased or noisy opinions, but it is an empirical question whether extreme opinions reflect signal, noise, or bias. I do find that extreme opinions contain more nationalistic bias on average, but the signal-to-bias and signal-to-noise ratios of these opinions do not decline until they become 2–3 times more extreme than the point at which they are normally truncated in ski jumping and 8–10 times more extreme than the normal truncation point in figure skating. Although judges' responses to reduced truncation may increase the bias content of extreme opinions, these results suggest that a modest reduction in truncation would improve decisions without increasing bias. In addition, truncating opinions into votes can increase the sustainability of vote trading agreements, for reasons I discuss below.

In organizations, it can be difficult to distinguish favoritism from tastes, and this is an issue in sports judging as well. For example, Basset and Persky (1995) and Campbell and Galbraith (1996) have documented systematically higher scores being given by judges from the same country in earlier Olympic figure skating competitions, but noted that these could be due either to a nationalistic bias or a taste for a particular national style of skating. There is also a question about whether biases are conscious or unconscious (e.g., arising from the fact that judges from a particular country may be more familiar with or more likely to identify with a particular athlete). Some of the cross-sectional variation in bias that I describe above is difficult to rationalize with tastes or unconscious biases, however. Examples include the fact that the national identity and past judging bias record of the other panel members appears to affect scores or the fact that biases vary in a way that accords with judges' career concerns.

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 et al., 2005) or that sumo wrestlers throw matches to each other

[4] Since the end of my sample period and in response to the judging scandal in the 2002 Olympics, the International Skating Union has modified its scoring and judge selection system. I will discuss these modifications in Section 7.

in response to nonlinearities in incentives (Duggan and Levitt, 2002). 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 et al. (2002) examine racial integration in baseball to determine whether leaders or followers are more likely to innovate, and Romer (2002) documents an excessive conservatism by American football coaches on fourth down and postulates organizational and behavioral explanations.

There is also an extensive theoretical literature on the problems of relying on information from potentially interested parties. The judging setting analyzed in the paper is most closely analogous to that of Prendergast and Topel (1996), who examine the problem of relying on the opinion about employee performance from one potentially biased supervisor.[5] 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 experts.[6] While this literature primarily focuses on the career concern of appearing to have high-quality signals, the empirical evidence in this paper suggests that the career concern of appearing unbiased (or, in the case of figure skating, biased) is more important for Olympic judges.

The remainder of the paper is divided as follows. Section 2 describes the data, and Section 3 describes the methodology and my estimate of average nationalistic bias. Section 4 examines cross-sectional variation in judging biases. As discussed above, I find that biases vary in ways that can be rationalized given judge's career concerns, and thus Section 5 examines the effect of past judging biases on the likelihood of achieving a judge's primary reward, the honor of being chosen to judge in the Olympics. Section 6 examines the information content of extreme opinions and argues that both figure skating and ski jumping may be truncating extreme opinions too aggressively. Two concluding discussions follow, one on the post-2002 proposals for judging reform and one on the possible implication of these results for decisionmaking in a corporate setting.

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

[6] For theoretical work on the career concerns of opinion producers and incentives for herding, see Scharfstein and Stein (1990), Trueman (1994), Brandenburger and Polak (1996), Ehrbeck and Waldmann (1996), Prendergast and Stole (1996), Avery and Chevalier (1999), and Laster et al. (1999). For empirical work on this same subject, see Lamont (1995), Chevalier and Ellison (1997, 1999), Hong et al. (2000), and Zitzewitz (2001).

Table I.
Sample Size

	Sport	
	Figure Skating	Ski Jumping
Events	16	23
Separate competitions (e.g., men's, women's)	61	25
Olympics	4	3
Pre-Olympics, Olympic-level	41	17
Post-Olympics, Olympic-level	0	5
Junior level	16	0
Rounds/Jumps	181	62
Compulsory (skating)/qualifying (skiing)	59	12
Short/long programs (skating)/finals (skiing)	122	50
Performances	2,976	2,920
Scores	25,068	14,600
Unique athlete countries	54	28
Unique athletes	584	243
Performances per athlete	5.1	12.0
Scores per performance	8.4	5.0
Unique judge countries	41	15
Unique judges	314	75
Events per judge	2.0	1.7

2. Data

The data are individual judge's scorings of figure skating and ski jumping 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 web site or elsewhere on the web.[7] For ski jumping, the sample includes the Olympics, all World Cup events, and the World Junior Championships. These events were included because score sheets, as opposed to just results, were available on the International Skiing Federation web site.

Table I summarizes the dimensionality of the dataset. It includes data on 16 figure skating and 23 ski jumping events. Essentially, all figure skating events include separate competitions for men, women, pairs,

[7] The ISU and other skating organizers use a software program called IceCalc to tabulate results and generate score sheets. Score sheets on web sites 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 II.
Summary Statistics for Scores

Sport	Score Type	Number of Scores Performance	Range	Mininum 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

Notes: In figure skating, each of 5, 7, or 9 judges issues scores for both technical merit and artistic impression. In ski jumping, 5 judges evaluate the style of a jump.

and ice dancing. The Olympic ski jumping competition included events on both 90 and 120 m hills, but World Cup events all involved only one of the two hill sizes. About half of the figure skating competitions include compulsory rounds, and similarly about half of ski jumping 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. Across all these events, I have data on close to 3,000 athletic performances in each sport.

Table II presents summary statistics for the judges' scores. In figure skating each of five, seven, or, usually, nine 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. I analyze both the judges' ordinal ranking and the sum of the scores, although the results are understandably very similar.[8] In ski jumping, each jump is scored unidimensionally on style by five different judges. Subjectively assessed style is combined with objectively measured distance to yield a total score for each jump. Distance accounts

[8] In both sports, I focus on points rather than placements as my primary metric. This choice increases the relative weight given to biases at the top of the standings, because, in both sports, the point difference between positions is 2–3 times greater at the top of the standings as it is further down (place 20 to 30). One might feel that a bias that affects first and second place is in some sense more important than the bias that affects 21st and 22nd, so emphasizing the measure that places greater weight on the first biases is probably appropriate.

for a large share of the variance in results: the standard deviation of distance points, style points, and total points is 20.9, 3.3, and 22.9, respectively.[9]

The aggregation of scores differs in the two sports. In ski jumping, the total style score is the mean of the middle three of five scores. Figure skating essentially truncates scores into votes about which skater performed better. The only aspect of a judge's scores that is influential is her relative ranking of any two competitors; any information contained in the difference in the scores assigned to the two skaters does not affect the results of the competition. The scores given each performance for technical merit (TM) and artistic impression (AI) are added together and then, based on their sum, 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.[10] 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 compulsories but less than the long program.

3. Measuring Nationalistic Biases

The primary empirical problem in measuring nationalistic biases is that one does not observe an objective measure of performance quality. One 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, objective measurements of the distance and speed of the jump.

Suppose that judges observe the objective quality of a performance with some error. They have a direct preference over the score given the event, which I call bias, and an ethical preference for scoring the performance as close to their subjective belief about its quality. Further suppose that the contribution to judge *j*'s utility function from scoring performance *p* by athlete *i* is

$$u_{ijp} = b_{ij} \cdot s_{ijp} - \frac{1}{2}(s_{ijp} - \hat{q}_{ijp})^2, \tag{1}$$

[9] Style point differences do occassionally have significant effects on athletes' placement, however. For example, Sven Hannawald of Germany dropped from second to fourth place in the 120 m hill competition at the 2002 Olympics entirely due to a poor style point score on his second jump. Consistent with the overall results in this paper, the German judge rated his second jump 13.5; the other four judges all scored it 11.5 or 12.0.

[10] This leads to occasional nontransitive 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.

where b_{ij} is the intensity of her direct preference (relative to her ethical concern) and $\hat{q}_{ijp}$ is her subjective assessment of quality. Given these preferences, the judge will score the performance as

$$s_{ijp} = q_{ip} + b_{ij} + e_{ijp}, \tag{2}$$

where s_{ijp} is the score, q_{ip} is the objective quality of the performance, b_{ij} again is the bias the judge has in favor of athlete i, and $e_{ijp} = \hat{q}_{ijp} - q_{ijp}$ can be viewed as the judge's observational error.

In the empirical work that follows, I assume that e_{ijp} is zero in expectation. Lacking an objective measure of performance, this assumption defines performance to be the average score given to the performance by the judges, after adjusting for any nationalistic bias in their scores. In order to separate bias from observational error and make this adjustment, I will assume that bias varies only with the judge–athlete country combination. If a judge from country J is especially biased against a specific performance or a particular athlete from country I, I will not capture this bias in my estimates. I also need to assume that the observational error e_{ijp} is zero in expectation for any judge–athlete country pair. If a judge has a unconscious preference for skaters from a given country and does not back that preference out of her scores, I will label this as bias. Bias, thus defined, might include tastes, but as discussed above, the apparent strategic variation suggests that it has a non-taste related component.

Given this setup, I use two different strategies to identify nationalistic biases. The first approach is to compare scores for the same performance given by different judges; the second uses athlete fixed effects and other observables to control for performance quality and assumes that the within-athlete variation in performance quality is uncorrelated with judges' 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 one is willing to assume that compensating biases are small, in other words that country I judge's evaluation of an athlete from country $J \neq I$ does not vary with whether country J is represented on the judging panel, then one can view this difference as being an approximation of the nationalistic bias.

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

$$s_{ijp} = B \cdot \Phi(I = J) + q_{ip} + l_j + e_{ijp}, \tag{3}$$

Table III.
Nationalistic Bias by Sport

Sport	Performances	Scores	Estimated Bias in Points		In Standard Deviations	
			Coeff.	SE	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

Notes: 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 IV, Line 1). Asterisks indicate significance at the 5% level (two-tailed). Stardard errors are heteroskedasticity-robust and allow for clustering within judge–athlete combinations.

where B is the average nationalistic bias, I and J index athlete and judge countries, $\Phi()$ is the indicator function, and q_{ip} is a performance *fixed* effect. A judge fixed effect, l_j, is added to capture any differences in "leniency" across judges. Nationalistic bias is measured as the average difference between the own-country score and the other judges' scores, the identifying assumption being that the latter are unbiased. Table III reports results from this method for the two sports. Biases are statistically and arguably also economically significant, particularly given their size relative to the within-performance standard deviation of scores.

Table IV uses the second identification approach, that allows me to relax the assumption of zero compensating biases. The second approach involves estimating the model:

$$s_{ijp} = B_{\text{nat}} \cdot \Phi(I = J) + B_{\text{comp}} \cdot \Phi(I \in P, I \neq J) + a_i + \beta x_{ip} + \tilde{q}_{ip} + l_j + e_{ijp}, \tag{4}$$

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 $\tilde{q}_{ip}$ 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

TABLE IV.
ALTERNATIVE SPECIFICATIONS AND IDENTIFICATION APPROACHES

Panel A. Ski Jumping

	Performance		Jump		Same Country		Different Country	
Line	Effects	Skier FEs	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*	(0.040)	−0.034*	(0.015)
(4b)	Random	Yes	Yes	14, 600	0.118*	(0.038)	−0.019	(0.017)
(5a)	Clustering	Yes	No	14, 600	0.105*	(0.039)	−0.034*	(0.015)
(5b)	Clustering	Yes	Yes	14, 600	0.118*	(0.031)	−0.019	(0.017)

	Performance		Objective		Same Country		Different Country	
Line	Effects	Sketer FEs	Characteristics	Obs.	Coeff.	S.E.	Coeff.	S.E.
Panel B. Figure Skating (TM + AI)								
(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)
(5)	Clustering	Yes	N/A	25, 068	0.256*	(0.025)	0.094*	(0.024)
Panel C. Figure Skating (Ordinal Placement)								
(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)
(5)	Clustering	Yes	N/A	25, 068	−2.552*	(0.222)	−1.797*	(0.223)

Notes: All regressions include judge country fixed effects. Regressions do not include performance fixed effects include fixed effects for event*round combinations. Regressions that control for jump characteristics include distance jumped and take-off speed and interactions of these variables with event fixed effects. Specifications labeled "clustering" include neither performance fixed, nor performance random effects, but have standard errors that adjust for within-performance clustering (Moulton, 1990). Standard errors are heteroskedasticity-robust. Asterisks indicate significance at the 5% level (two-tailed) Dependent variable: style points (ski jumping) or TM+AI score total (figure skating).

judging panel. In other words, the key assumption is that a given athlete does not perform better when certain countries are represented on the judging panel, except to the extent that these differences are captured by observables.[11]

For figure skating, the x_{ip} includes fixed effects for 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 event-round combination. For ski jumping, the x_{ip} also includes the distance jumped and take-off speed for each jump. All else being equal, a ski jumper with the same take-off speed but better form while airborne will travel further. For example, when I regress style points on distance and speed, I find positive and negative coefficients, respectively, with the regression explaining about 50% of the variation in style points. If I loosely call the style that is statistically explained by distance and speed "airborne" style and the residual "landing" style, then including jump characteristics relaxes the 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 IV suggest that measuring $B_{\text{nat}} - B_{\text{comp}}$ is much easier than separately measuring its components, but I can obtain statistically significant estimates of B_{comp} for both sports. 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 athlete and event-round combination fixed effects yields a point estimate of -0.033 (SE $= 0.045$); a similar regression for the median score in figure skating yields an estimate of 0.116 (SE $=$ 0.025), which implies that having a compatriot on the judging panel is very important.[12] The positive

[11] One situation where one might expect athlete performance to be correlated with judging panel composition is for host-country athletes, who may both perform better at home and be more likely to have a compatriot judge on the panel. As a robustness check, I reestimated all results omitting host country athletes and did not find that they were qualitatively different.

[12] One might suppose that given that most figure skating events have nine judges, the top countries are always represented, and so the representation advantage only affects the bottom of ths standings. This actually is not the case; in my sample, athletes from the top four countries in terms of rounds won (Russia, the US, Canada, and France) had only a 70% chance of being represented in a given event. Thus the panel representation advantage affects results at the top, as well as the bottom, of the standings.

B_{comp} in figure skating and negative B_{comp} in ski jumping also implies that comparing the scores awarded a given performance (the first identification approach) overstates nationalistic biases in ski jumping while understating them in figure skating.

4. Variations in Nationalistic Biases

Table V examines variation in nationalistic bias by estimating it for subsamples of the data. When examining bias in subsamples, I am forced to use the first identification approach due to sample size constraints.

In ski jumping, nationalistic biases are larger 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 90 m hill, where style points account for a large amount of the variance in total scores. Nationalistic biases do exist even for the scored qualifying jumps of pre-qualified skiers, despite the fact that these jumps have no effect on the competition.[13]

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 (albeit not significantly), 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 compulsories, 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 these differences is that whereas the scrutiny of style judging in ski jumping may be limited, in 2001–2002 scrutiny was fairly significant in figure skating and was probably especially so for more important competitions and for noncompulsory rounds that are watched by larger audiences. In terms of the simple model above, b is the ratio of the judge's preference over the results and her concern for accuracy. Preferences may be stronger for more

[13] Two possible reasons why judges may be biased in favor of their countrymen even when doing so has no effect on the results are (1) judges are not fully conscious of their nationalistic biases, or (2) judges attempt to maintain consistency in their scoring of a given athlete.

Table V.
Nationalistic Biases in Subsamples of the Data

		Nationalistic Bias		p-Value Of Bias Difference with Next Category
	Obs.	Coeff.	S.E.	
Panel A. Ski Jumping				
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)				
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: This table replicates the regression in Table IV, Line 1 for subsamples of the data. To avoid a sample selection bias, in which athletes are "top 10" or "not top 10" is determined by replacing the score in the current observation with the average other score given to that performance and then reranking the competitors in that contest. Asterisks indicate that the average nationalistic bias for a given subsample of the data is statistically significantly different from zero (at the 5% level, one-tailed); the p-values are for the significance of the difference between two categories.

important events, but increased scrutiny may also increase the concern for accuracy. The results in Table V suggest that the former dominates for ski jumping while the latter dominates for figure skating.

Another source of variation in nationalistic bias is by country. Table VI presents estimates of country-specific nationalistic bias estimated using the following model, which is a version of (3),

$$s_{ij} = B_J \cdot \Phi(I = J) + L_J + q_{ip} + e_{ijp}. \tag{5}$$

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 (3). One might note that nationalistic biases appear to rise with eastern longitude, and they also appear to rise with indices of perceived country-level corruption, such as the *Transparency International* index included in the Table VI.[14]

If a desire for fairness explains the negative compensating bias in ski jumping, judges should compensate more against athletes from countries with a history of being more nationalistic. Table VII presents results from estimating a version of (4) that allows B_{comp} to vary with the nationalistic bias of the athlete's representative on the judging panel,

$$\begin{aligned} s_{ij} = {} & B_{\text{same}} \cdot \Phi(I = J) + (C_{\text{comp}} + D_{\text{comp}} \cdot B_I) \cdot \Phi(I \in P, I \neq J) \\ & + a_i + \beta x_{ip} + \tilde{q}_{ip} + e_{ijp}. \end{aligned} \tag{6}$$

B_{comp}, the bias by other country judges toward country I athletes when I is represented on the judging panel, is allowed to vary linearly with the nationalistic bias of judges from country I. A negative D_{comp} would indicate a desire for fairness, while a positive C_{comp} would indicate a "panel representation effect," or a bias in favor of athletes represented by judges from unbiased countries. The results in Table VII 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_{comp}, 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

[14] The 2001 index is used to avoid using an index that might have been influenced by the judging scandal at the 2002 Olympics. The simple cross-country correlation between the index of perceived noncorruption and judging nationalism is −0.59 for ski jumping and −0.38 for figure skating.

Table VI.
Nationalistic Bias and Leniency by Judge Country

		Nationalistic Bias		Leniency		Observations		Transparency International
						All	Same-Country	Corruption Perceptions
Country	Abrev.	Coeff.	SE	Coeff.	SE	Scores	Athlete	Index (2001)
Panel A. Ski Jumping								
South Korea	KOR	0.386*	(0.203)	0.175*	(0.049)	80	4	4.2
Slovakia	SVK	0.297*	(0.121)	0.071*	(0.025)	617	11	3.7
France	FRA	0.292*	(0.111)	0.017	(0.026)	583	13	6.7
Czech Republic	CZE	0.255*	(0.082)	0.000	(0.021)	467	25	3.9
Slovenia	SLO	0.249*	(0.041)	0.002	(0.022)	1,279	104	5.2
Sweden	SWE	0.246*	(0.230)	−0.101*	(0.025)	610	3	9
Germany	GER	0.180*	(0.027)	0.079*	(0.020)	2,046	245	7.4
Austria	AUT	0.153*	(0.027)	0.103*	(0.020)	2,246	253	7.8
Poland	POL	0.145*	(0.057)	0.091*	(0.026)	555	53	4.1
Italy	ITA	0.144	(0.100)	0.119*	(0.024)	762	16	5.5
Norway	NOR	0.109*	(0.033)	−0.093*	(0.021)	1,839	158	8.6
Finland	FIN	0.081*	(0.034)	−0.035	(0.023)	1,256	153	9.9
Japan	JPN	0.041	(0.035)	0.060*	(0.024)	900	159	7.1
Switzerland	SUI	0.024	(0.077)	0.008	(0.025)	542	28	8.4
USA	USA	0.008	(0.078)	0.014	(0.023)	818	27	7.6
Panel B. Figure Skating								
Azerbaijan	AZE	0.316*	(0.075)	0.001	(0.051)	699	28	2.0
Hungary	HUN	0.310*	(0.068)	−0.064	(0.051)	826	34	5.3
Slovenia	SLO	0.306*	(0.113)	0.057	(0.052)	480	12	5.2
Romania	ROM	0.300*	(0.109)	0.024	(0.052)	542	13	2.8
South Korea	KOR	0.290*	(0.124)	0.018	(0.055)	258	10	4.2
Slovakia	SVK	0.248*	(0.079)	−0.015	(0.051)	739	25	3.7
Uzbekistan	UZB	0.231*	(0.095)	0.016	(0.055)	260	18	2.7
Poland	POL	0.225*	(0.049)	−0.033	(0.050)	929	66	4.1
Canada	CAN	0.210*	(0.032)	−0.108*	(0.050)	1,610	158	8.9
Italy	ITA	0.205*	(0.046)	−0.083	(0.050)	1,167	76	5.5
Czech Republic	CZE	0.174*	(0.058)	−0.029	(0.051)	889	47	3.9
USA	USA	0.172*	(0.033)	−0.087	(0.050)	1,387	159	7.6
Belgium	BEL	0.156	(0.124)	−0.064	(0.053)	392	10	6.6
Germany	GER	0.155*	(0.045)	−0.016	(0.050)	1,423	80	7.4
Finland	FIN	0.144*	(0.063)	−0.055	(0.051)	790	39	9.9
China	CHN	0.132*	(0.057)	0.082	(0.053)	454	52	3.5
Australia	AUS	0.132*	(0.051)	−0.076	(0.051)	1,077	60	8.5
Japan	JPN	0.129*	(0.044)	0.037	(0.050)	1,238	83	7.1
Russia	RUS	0.117*	(0.030)	0.063	(0.050)	1,603	190	2.3
Bulgaria	BUL	0.115	(0.079)	−0.028	(0.051)	698	25	3.9

Continued

TABLEVI.
CONTINUED

		Nationalistic Bias		Leniency		Observations		Transparency International Corruption Perceptions Index (2001)
Country	Abrev.	Coeff.	SE	Coeff.	SE	All Scores	Same-Country Athlete	
Panel B. Figure Skating								
France	FRA	0.114*	(0.038)	−0.037	(0.050)	1,156	114	6.7
Switzerland	SUI	0.114*	(0.060)	−0.017	(0.050)	1,096	43	8.4
Ukraine	UKR	0.110*	(0.040)	0.000	(0.050)	1,331	101	2.1
Estonia	EST	0.063	(0.074)	0.002	(0.052)	575	29	5.6
Great Britain	GBR	0.041	(0.074)	−0.238*	(0.051)	702	29	8.3

Note: 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.

*indicate significance at the 5% level (one-tailed for bias; two-tailed for leniency).

TABLE VII.
VARIATION OF COMPENSATING BIAS WITH THE NATIONALISTIC BIAS OF OTHER JUDGES

	Ski Jumping		Figure Skating	
	Coeff.	SE	Coeff.	SE
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: Dependent variable: style point score (ski jumping) or TM + AI score total (figure skating). 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 IV).

judge is more than compensated for in each of the other judges' scores.[15]

[15] Some readers have wondered whether there is a mechanical relationship that forced D_{comp} to be negative. Equation (6) takes the second identification approach, so bias is identified using athlete rather than performance fixed effects. The compensating bias of judges from country J against country I athletes is thus estimated by comparing how the scores awared to specific athletes from country I by judges from J vary with whether I is represented on the panel. D_{comp} thus picks up whether the compensating bias against

The fact that D_{comp} is negative in both sports is inconsistent with one potential explanation for the positive compensating bias in figure skating: "yes men"-style herding to avoid reputation-harming disagreements (Prendergast, 1993). 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 one would expect D_{comp} to be positive, that is, 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, I can test for whether biases are reciprocated. To do this, I estimate a version of (4),

$$s_{ij} = B_{IJ} + a_i + \beta x_{ip} + q_{ip} + e_{ijp}, \tag{7}$$

where B_{IJ} is the average bias of a judge from country J in favor of an athlete from country I. I can then perform a simple test for reciprocity by examining the correlation between $\hat{B}_{IJ}$ and $\hat{B}_{JI}$. In figure skating, the $\hat{B}_{IJ}$ and $\hat{B}_{JI}$ 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, I also test for whether bloc voting can help explain the patterns in the $\hat{B}_{IJ}$. I estimate by maximum likelihood a model in which there are two voting blocs, and $B_{IJ} = B_{\text{same}}$ if I and J are members of the same bloc or $B_{IJ} = B_{\text{diff}}$ if they are different. I allow the nationalistic biases, that is, the B_{II}, to vary freely for each country. In figure skating, the likelihood is maximized for the top 10 countries by a model in which the US, 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 nonaligned (Table VIII).[16] The estimate $\hat{B}_{\text{same}} = 0.001$ and $\hat{B}_{\text{diff}} = -0.051$, so together with Table II 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. An F-test suggests that the version of (7) allowing for voting blocs describes the figure skating data significantly better than a model with only a nationalistic and compensating biases (i.e., equation (7)).

I (the difference between a country I athlete's scores from $J \neq I$ when I is on the panel and when I is not) is related to I's nationalism (the difference between country I athlete's scores from judges from I and judges from $J \neq I$).

[16] The results exclude the top two couples in the pairs competition at the 2002 Olympics, in which the media speculated there had been bloc voting along roughly these lines. Bloc voting along Cold War lines in the 1968–1988 Olympics was documented by Seltzer and Glass (1991).

Table VIII.
Bias Matrix for Figure Skating

Panel A. Bias Matrix

Athlete	Judge Country									
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

Note: 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.

Unsurprisingly, given the lack of reciprocity in ski jumping noted above, a voting bloc model does not describe the ski jumping data better than the model in (7).

5. Career Concerns of Judges

One possible motivation for the concern for accuracy in the simple model above is a judge's career concerns. In an organizational setting, a committee member would like to avoid the appearance that her opinions are biased or noisy in order to achieve greater future influence. In sports judging, future influence comes in the form of being chosen to judge important competitions such as the Olympics. Given that most of my sample for both sports consists of events immediately before 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.

Table IX 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. It 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 I 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*. Because the selection process for judges is often two-stage, with countries to be represented chosen first and representatives chosen second, I 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.

The most important distinction of the ski jumping judges who were selected for the Olympics was that they displayed essentially no nationalistic bias.[17] For figure skating, however, the judges chosen for the Olympics are both statistically significantly more lenient and *more* nationalistic.

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 as 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. Because 20 different national federations send judges to the Olympics, it is quite plausible that cooperation on selecting unbiased judges is difficult to sustain.

Some of the cross-sectional variation in nationalistic bias discussed in Table V can be rationalized given career concerns. Ski jumping judges

[17] This difference is statistically significant, but the sample size of Olympic ski jumping judges is admitedly small: six ski jumping judges judged in the 2002 Olympics, but only three judged the 2001–2002 World Cup events that were in our sample. Probit regressions predicting selection to judge in the Olympics confirm the statistical significance of the relationships discussed above for both sports.

TABLE IX.
CHARACTERISTICS OF JUDGES CHOSEN TO JUDGE IN THE OLYMPICS

	(1) Chosen for Olympics		(2) Not Chosen		(3) Not Chosen and from Country with Judge		*P*-Values for Comparing Means	
	Mean	SD	Mean	SD	Mean	SD	(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

Note: 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). The *p*-values reported are heteroskedasticity robust.

are more nationalistically biased in the Olympics; this is consistent with their surpressing their biases in order to achieve future influence. In contrast, figure skating judges are actually less nationalistically biased at the Olympics. While this may simply be due to the extra scrutiny, another possibility is that judges are actually more nationalistic than they would like to be in pre-Olympic events in order to curry favor with the national federations. Once they are in the Olympics, their concern for fairness reasserts itself, although apparently only partially.

6. Nonlinear Aggregation of Scores

Both ski jumping and figure skating underweight extreme opinions. In ski jumping, a skier's total style score is the sum of the middle three of five scores. If an individual judge's scoring of a jump 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 nontransitivities as discussed above).

Whether underweighting extreme opinions is statistically optimal *ex post* depends on whether the signal-to-noise and signal-to-bias ratios decline as they become more extreme. Whether it is optimal as an *ex ante* mechanism also depends on the incentives such a mechanism creates for judges. This section will begin by empirically examining whether extreme truncation is statistically optimal, and then turn to the issue of the incentives it creates for judges.

A principal interested in a one-shot minimum mean-squared error evaluation of a performance would simply evaluate the performance using $E(q \mid S)$. For example, if a score is the sum of objective quality, bias, and observational error $s = q + b + e$ as in (2) above and each component is independently and normally distributed, then $E(q \mid S)$ will be linear in s. If the bias and observational errors are independent across a set of scores S, then $E(q \mid S)$ will be linear in the scores.

Extreme truncation is only statistically optimal in cases where the signal-to-noise and signal-to-bias ratios of opinions decline as they become more extreme. How can one determine whether this is the case? If one observed an objective measure of performance quality, one could examine the predictive power of extreme opinions. In other words, one could ask: when one judge's opinion is very different from that of the others, how much weight should one be putting on the extreme opinion in constructing an estimate of performance quality? If one takes the mean of the opinions of $J - 1$ judges as a starting point, how should one revise one's expectation of quality based on the difference between the *J*th opinion? One would like to estimate the function $f()$ in

$$E(q \mid S) - \bar{s}_{-j} = f(s_j - \bar{s}_{-j}), \tag{8}$$

where $\bar{s}_{-j}$ is the average of all scores other than judge *j*'s.

If one observed q, one could estimate this function using techniques similar to those used by Zitzewitz (2001) to analyze analyst forecasts. Specifically, by nonparametrically estimating:

$$\begin{aligned} q - \bar{s}_{-j} &= f(s_j - \bar{s}_{-j}) + \varepsilon_j, \\ \varepsilon_j &= [q - E(q \mid S)]. \end{aligned} \tag{9}$$

This estimate will be consistent if the expectation of the error term is zero for all values of the right-hand side variable, which must be true since $s_j - \bar{s}_{-j}$ is included in S.

An important difference between the judging data and the analyst data is that in the judging data one does not observe an objective measure of q, so one cannot estimate (8). What one can do instead is take the opinion of judge k to be the objective measure and then study what method of aggregating the other opinions leads to the best estimate of k. Specifically, one would nonparameterically estimate the equation

$$\begin{aligned} s_k - \bar{s}_{-jk} &= f(s_j - \bar{s}_{-jk}) + \varepsilon_{jk}, \\ \varepsilon_{jk} &= (s_k - q) + [q - E(q \mid S_{-k})], \end{aligned} \tag{10}$$

Again, for estimation to be consistent, $E(\varepsilon_{jk} \mid s_j - \bar{s}_{-jk})$ must be zero. This is clearly satisfied for the second term, so the necessary assumption reduces to $E(s_k - q \mid s_j - \bar{s}_{-jk}) = 0$. If we write $d_k = s_k - q = e_k + b_k$ for judge *k*'s sum of error and bias, then this assumption can be rewritten as $E(d_k \mid d_j - \bar{d}_{-jk}) = 0$. In other words, judge *k*'s deviation from the truth should be independent of the difference between judge j and all her other colleagues.

An easy way of ensuring that d_k is *uncorrelated* with $d_j - \bar{d}_{-jk}$ 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 d_k not be related to higher moments of the distribution of $d_j - \bar{d}_{-jk}$. Given the finding in this section that there is a surprising amount of information in extreme opinions, what one should worry most about is d_k being positively related to the skew in the $d_j - \bar{d}_{-jk}$. 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

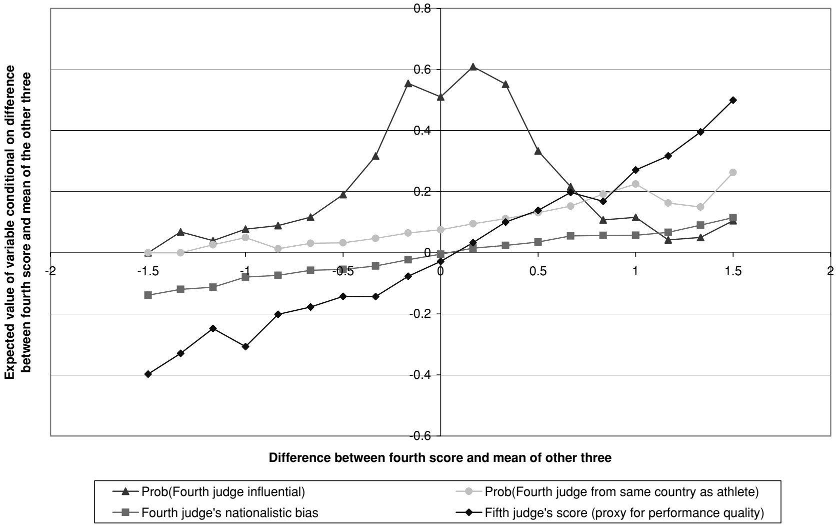


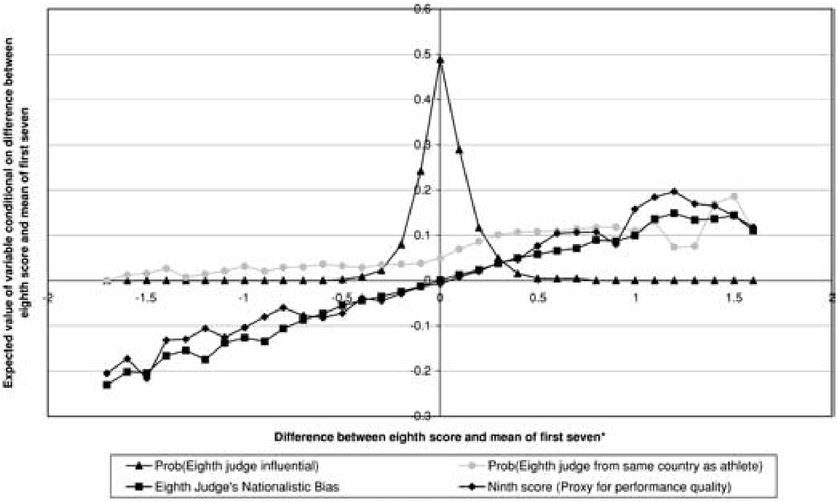
FIGURE 1. SIGNAL AND BIAS CONTENT OF EXTREME JUDGE OPINIONS—SKI JUMPING

the likelihood of making them for a given performance were correlated. In what follows, I make the necessary independence assumption.[18]

The slope of the 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, I also estimate the function $g(s_j - \bar{s}_{-jk}) = E(B_{IJ} \mid s_j - \bar{s}_{-jk})$, where the B_{IJ} are those estimated for the athlete-judge country combination in (7). This captures only the nationalistic and cross-national 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, I 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. I also graph the probability of a score being influential. A score is

[18] The working paper version of this paper (Zitzewitz, 2002) examines the robustness of the conclusions of this section to this assumption. In particular, it works through an example designed to maximize the correlation between d_k and the skewness in $d_j - \bar{d}_{-jk}$ and finds that the conclusions drawn from using the method proposed here differ only slightly from what one would conclude if one observed objective performance quality and thus could estimate (8) directly. In the example analyzed, all of the excess kurtosis in the observed distribution of figure skating scores is assumed to come from the source that is most problematic for the method in question, that is, from a common, large-variance, observational error that affects exactly two out of nine judges.



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

FIGURE 2. SIGNAL AND BIAS CONTENT OF EXTREME JUDGE OPINIONS—FIGURE JUMPING

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% of the sample in each sport. Given the aggregation 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 one might also expect the bias to increase due to the incentives created for judges. But these results suggest that at least small movement in this direction may reduce both the noise and bias content of performance evaluations.

6.1 Extreme Truncation and Vote Trading

Two interesting cross-sport correlations emerge from the winter sports judging results. First, the sport that engages in more truncation of extreme opinions (figure skating) has the most nationalistically biased judging results. In addition, the three judged winter sports omitted from this version of the paper (moguls, aerials, and snowboarding) all had less nationalistic bias and did less truncation of extreme scores than either ski jumping or figure skating (see Zitzewitz, 2002 for details). Second, figure skating, the sport with extreme truncation, displays evidence of vote trading, while in ski jumping 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 $q + e_j$, where q is the true difference between the quality of two performances and e_j is an observational error. Suppose that in one system, judges report $s_j = q + e_j + b_j$, and the evaluation is the average of these reports across judges, whereas in another system, the judges report only the sign of s_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: $q + e_j < -b_j$. In the first system, the principal observes a less noisy signal about the b_j used by the judge, $s_j - \hat{q}$, instead of simply observing whether b_j was greater than of less than $-(q + e_j)$, so updating prior beliefs about judge's biases is faster, and thus the reputational 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.

Finally, vote trading agreements are less flexible than reciprocal bias agreements with continuous scores. Colluding judges must decide in real time whether a performance was bad enough to justify deviating from an agreement; they have no meaningful way of simultaneously acknowledging both the performance quality and their fidelity to the collusive arrangement.[19] This lack of flexibility could make vote trading arrangements hard to sustain, but it should cause those arrangements that are sustained to determine outcomes for a greater range of actual performance qualities.

7. Implications for Judging Reform

These results are potentially interesting at two levels. First, in light of the number of people who watch the Olympics and their nontrivial economic size, the fairness of judging at the Olympics and other sporting events is presumably of direct interest. Second, organizations often make committee decisions that are similar to Olympic judging. Institutional features such as the selection of committee members and the aggregation of opinions affect the quality of decisions in the Olympics, and it is not inconceivable that they would have analogous effects in organizations. The final two sections of the paper discuss the implications of the results for the reform of Olympic judging and for the design of group decision-making processes more generally.

Following the figure skating judging scandal at the 2002 Olympics, the International Skating Union (ISU) modified its judging system in two stages. In a June 2002 meeting, the ISU adopted a proposal to have it, rather than the national federations, choose judges, as the FIS does in skiing. In addition, it adopted a proposal to have 14 rather than 9 judges judge each event. Under this new system, all 14 scores would be publicly reported, but only 9 would be used in computing results, and which scores counted and which judge gave which score would not be publicly revealed. Subsequently, it adopted two more reforms: (1) to replace the technical merit and artistic impression scores with a system such as the one used in gymnastics or diving, where elements of a routine are assigned predetermined "degrees of difficulty" and judges graded the quality execution of these elements, and (2) to aggregate scores by

[19] In some cases, colluding judges attempt to get around this problem by communicating in real time about whether their vote trading aggrement still applies given the quality of the performance. For example, on April 28, 2002, *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.

Table X.
Implications of the Results of the Paper for Current Proposals for Reforming Figure Skating Judging

	Implications of Paper's Findings for Desirability	
Proposal and Aspect	Sign	Rationale
June 2002 reforms		
ISU selects individual judges, rather than national federations	+	Career concerns results suggest that FIS chooses less biased judges in ski jumping, while figure skating national federations choose more biased judges.
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	0	In most cases, observers should be able to determine which scores were used from the aggregated results
Do not reveal which judge gave which scores	?	Should reduce collusive agreements harder by making defection from them easier, but one would then need to trust the ISU to monitor judges for bias.
Subsequent reforms (2003 and 2004)		
Technical merit and artistic impression scores replaced with a more objective degree of difficulty measure, modified by judges' evaluation of execution	Probably +	Should reduce subjective component of scoring, reducing scope for bias, but some athletes argue that lower weight on artistry will reduce sport's appeal
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.

taking the average of the middle five out of nine scores, similar to ski jumping's system of averaging the middle three of five scores.

The results of this paper yield some insights that allow us to comment on these proposals (Table X). First, the results judge's career concerns in Section 5 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 is also probably a positive change. Aerials, which uses this system, had the smallest estimated nationalistic bias, relative to the within-performance standard deviation, of the five winter sports studied by Zitzewitz (2002). The multiple controversies over the degrees of difficulties assigned to the gymnastics routines in the 2004 Olympics suggest though that even these allow for some subjectivity. In addition, some skaters have criticized the new system as devaluing the artistic component of skating, since this component is more difficult to score objectively than technical merit, these complaints may partly reflect a fundamental trade-off between a system's objectivity and the weight it can put on artistry.

Third, having 14 judges score the competition but only nine judge's scores count, is best considered in 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 nine of the 14 scores will simply be the results from aggregating all 14 scores plus noise (assuming that the randomization is fair, about which there may be skepticism). 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.

Although it is not clear what is accomplished by not revealing which nine 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 I find smaller biases when scrutiny is likely to be higher, a lack of monitoring by outsiders may lead an increase in nationalistic biases.[20] 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, cross-sport evidence suggests that less truncation is correlated with less biased results. Although this relationship is not necessarily causal, the analysis in Section 6 suggests that the truncation of opinions into votes involves substantial loss of information, and there are reasons to believe, as outlined in Section 6.1, that extreme truncation encourages vote trading. This suggests that the change from voting to the truncated mean of scores used in many other Olympic sports (ski jumping, moguls, and aerials, as well as gymnastics and diving) may be an improvement.

8. 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 I 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

[20] The United States Figure Skating Association has expressed precisely this concern with the anonymity of judges. The ISU has rejected calls to drop anonymity, however, because it argues that it makes judges more difficult to influence (see Shipley, 2003).

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 "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 observed in figure skating.

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

1. *Value 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, as opposed to an alternative norm of celebrating committee members' political skill in promoting their interests, it can help lead to unbiased decision-making even with biased decision-makers.
2. *Use 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. *Recognize the costs of opinion truncation*. The results in Section 6 suggest that truncating opinions into votes uses information inefficiently, and the discussion in Section 6.1 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.
4. *If all else fails, balance the committee*. Figure skating has done this since the controversial 1927 World Championships in which Sonia Henie of Norway was chosen first by the three Norweigian judges and the other judges, a German and an Austrian, voted for Herma Plank-Szabo of Germany. If bias cannot be otherwise constrained, set up the committee like a court room, with both sides represented (à la Milgrom and Roberts, 1986).

References

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

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

Avery, C.N. and J.A. Chevalier, 1999, "Herding Over the Career," *Economics Letters*, 63, 327–333.

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

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

Brandenburger, A. and B. Polak, 1996, "When Managers Cover Their Posteriors: Making Decisions the Market Wants to See," *RAND Journal of Economics*, 27, 523–541.

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

Campbell, B. and J.W. Galbraith, 1996, "Non-Parametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments," *The Statistician*, 45, 521–526.

Chevalier, J.A. and G.D. Ellison, 1997, "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, 105, 1167–1200.

——, and ——, 1999, "Career Concerns of Mutual Fund Managers," *Quarterly Journal of Economics*, 114, 389–432.

Duggan, M. and S.D. Levitt, 2002, "Winning Isn't Everything: Corruption in Sumo Wrestling," *American Economic Review*, 92, 1594–1605.

Ehrbeck, T. and R. Waldmann, 1996, "Why Are Professional Forecasts Biased? Agency Versus Behavioral Explanations," *Quarterly Journal of Economics*, 111, 21–40.

Garicano, L., I. Palacios, and C. Prendergast, 2005, "Favoritism Under Social Pressure," *Review of Economics and Statistics*, 87, 208–216.

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

Holmstrom, B., 1999, "Managerial Incentive Problems: A Dynamic Perspective," *Review of Economic Studies*, 66, 169–182.

Hong, H., J.D. Kubik, and A. Solomon, 2000, "Security Analysts' Career Concerns and Herding of Earnings Forecasts," *RAND Journal of Economics*, 31, 121–144.

Lamont, O., 1995, "Macroeconomic Forecasters and Microeconomic Forecasts, " NBER Working Paper 5284.

Laster, D., P. Bennett, and I.S. Geoum, 1999, "Rational Bias in Macroeconomic Forecasts," *Quarterly Journal of Economics*, 114, 293–318.

Meyer, M., P. Milgrom, and J. Roberts, 1992, "Organizational Prospects, Influence Costs, and Organizational Change," *Journal of Economics and Management Strategy*, 1, 9–35.

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

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

——, and L. Stole, 1996, "Impetuous Youngsters and Jaded Old-Timers: Acquiring a Reputation for Learning," *Journal of Political Economy*, 104, 1105–1134.

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

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

Scharfstein, D. and J. Stein, 1990, "Herd Behavior and Investment," *American Economic Review*, 80, 465–479.

Shipley, A., 2003, "ISU to Consider Changing Display of Scoring Marks." *The Washington Post,* March 30, D9.

Trueman, B., 1994, "Analyst Forecasts and Herding Behavior," *Review of Financial Studies*, 7, 97–124.

Zitzewitz, E., 2001, "Measuring Herding and Exaggeration by Equity Analysts." Stanford Univeristy Mimeo.