DISCUSSION PAPER SERIES

IZA DP No. 1257

Outlier Aversion in Evaluating Performance: Evidence from Figure Skating

Jungmin Lee

August 2004

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Outlier Aversion in Evaluating Performance: Evidence from Figure Skating

Jungmin Lee
University of Arkansas-Fayetteville
and IZA Bonn

Discussion Paper No. 1257
August 2004

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1257
August 2004

ABSTRACT

Outlier Aversion in Evaluating Performance: Evidence from Figure Skating*

The quality of subjective performance evaluation is dependent on the incentive structures faced by evaluators, in particular on how they are monitored and themselves evaluated. Figure skating competitions provide a unique opportunity to study subjective evaluation. This paper develops and tests a simple model of what I call "outlier aversion bias" in which subjective evaluators avoid submitting outlying judgments. We find significant evidence for the existence of outlier aversion. Individual judges within a game manipulate scores to achieve a targeted level of agreement with the other judges. Furthermore, a natural experiment shows that the dispersion of scores across judges depends upon the type of judge-assessment system and its implication for outlier aversion. Agreement may not be a good criterion for the validity of an evaluation system, contradicting the industrial psychology and personnel management literature.

JEL Classification: D7, M5

Keywords: subjective performance evaluation, outlier aversion bias

Jungmin Lee
Department of Economics
University of Arkansas-Fayetteville
Austin, TX 78712
USA
Email: JLee@walton.uark.edu

* I would like to thank Dan Hamermesh, Sokbae Lee, Gerald Oettinger, Max Stinchcombe, Steve Trejo, Justin Wolfers, Eric Zitzewitz, and participants at seminars at the University of Texas, the 11th International Panel Data Conference, and the 2004 North American Econometric Society summer meetings.

1 Introduction

Many important situations are judged by a subjective evaluation process. Examples include the evaluation of employees by their supervisors, firms by their customers and investors, academic articles by referees, and competitive athletes by panels of judges. In these cases, objective measures are either impractical or distorted, making subjective measures the only available choice.[1] However, it is well known that there are chronic problems with subjective performance appraisals. One inherent weakness is that these evaluations cannot be verified by any other than evaluators themselves. It is impossible to figure out the underlying processes by which evaluators reach their judgment. As a result, subjective measures can be possibly manipulated by the evaluators who are pursuing their own goals other than unbiased reviews. Accurate evaluation might be a relatively minor concern of assessors compared with their own rent seeking. Subjective evaluations are prone to bias.[2]

The quality of subjective evaluation, such as accuracy and unbiasedness, depends on the incentive system caused by the organization in which the evaluator is judging. The organization needs to devise an optimal mechanism in which subjective evaluators cannot manipulate their judgment in an arbitrary way. A simple, and perhaps the most popular, way of checking subjective evaluation is to employ multiple evaluators and to compile their opinions.[3] This may prevent individualistic bias, for example nepotic favoritism, since the organization can detect "unusual" evaluation by comparing different evaluations. Furthermore, aggregating multiple appraisals is supposed to average out individuals' heterogeneous preferences and idiosyncratic measurement errors. When different raters independently provide similar ratings for the same performance, it is accepted as a form of consensual validity or convergent validity [Murphy and Cleveland 1991, Saal, Downey, and Lahey 1980].

However, when the organization utilizes multi-source performance appraisals, evaluators have the incentive to eschew extreme opinions, thereby slanting the evaluations toward consensus. I call this tendency of gravity "outlier aversion bias."[4] Evaluators suppress the desire to question alternatives in favor of agreement seeking. They would conceal some information that they think others do not know. They would not want to be branded as "nonconformists" [Bernheim 1994]. A concurrence-seeking tendency, the so-called "groupthink," occurs because

groups desire unanimity and members are under considerable pressure to make a quality decision [Janis 1983][5]

This paper is most closely related to previous work by Prendergast [1993], who shows that when there are a supervising manager and subordinate evaluators, the evaluators have an incentive to conform to the opinion of the supervisor. In his model, the so-called "yes men" syndrome occurs when the subordinate evaluators' reports are compared with the manager's opinion that is based on his own observation and his observation on the opinions of subordinate evaluators. The analysis presented here differs from that of Prendergast in that I do not suppose any hierarchic relation, but assume that evaluators share some common knowledge. I show that when evaluators are appraised through comparison with their peers, there will be an incentive for all the evaluators to distort their assessments and to be biased toward a general consensus, which is often derived from assumptions based on pre-performance public information.

Unfortunately, there are few empirical studies on subjective performance evaluation.[6] Due to the lack of appropriate "economic" data, several studies have recently used sports data to show that evaluators in sports, such as referees, judges, or umpires, have many incentives to bias their decisions in the pursuit of objectives other than an accurate evaluation. There exists a striking degree of subjectivity in sports judgments. Garicano, Palacios, and Prendergast [2001] show that in soccer games, judges feel pressured by home-team spectators, who are usually a majority, and, as a result, manipulate players' injury time to promote the victory of the home team. Also, in the case of international figure-skating competitions, which this paper examines in further detail, it is found that there often exists patriotic favoritism toward skaters from the judge's native country [Campbell and Galbraith 1996, Zitzewitz 2002].

Using individual judges' scoring data from World Figure Skating Championships between 2001 and 2003, I examine the voting behavior of internationally-qualified judges and investigate whether they avoid submitting extreme scores. Figure skating is an excellent sport for testing "outlier aversion bias" because its judging process is almost entirely subjective. It provides a unique opportunity to test theories about subjective evaluation because we can repeatedly observe individual judges and their scores for different performances. Furthermore, a major Olympic scandal in 2002 involving the gold medal award for pairs skating caused a

discrete change in the organization and system of evaluation.[7]

The empirical analysis is based on the hypothesis that if outlier aversion exists and the judges' incentive structure for voting is altered because of it, the resulting scores and their distribution will likewise change in a systematic fashion. This paper exploits two kinds of variation in the judges' incentives: (i) endogenous variation of an individual judge's aversion to outlying scores over the course of performances according to the degree of extremity of his or her previous scorings within a game; and (ii) exogenous and across-the-board changes in all the judges' incentives for outlier aversion due to the judging system reform after the 2002 scandal. In both cases we expect to find a systematic change in voting behavior, which represents a similar change in the judges' incentive structure.

The judges aggressively avoid extreme scores. Once a judge has already submitted outlying scores for previous skaters in the competition, there is a strong tendency toward agreement with the other judges afterwards. We also find that after the launch of the new judging system that relieves judges of social pressures, the dispersion of scores across judges for the same performance significantly increased, even after controlling for the possibility that judgments are in less agreement because the reform aggravates nationalistic favoritism. Overall this study confirms the view that subjective evaluations are sensitive to the incentive provided by appraisal system. In particular, our findings imply that emphasizing concurrence in opinion would lead to significant loss of information on the margin by compressing evaluations around consensus.

The paper is organized as follows: In Section 2, we develop a simple behavioral model of strategic judgment and derive some testable hypotheses for outlier aversion bias. Section 3 describes the sample and presents empirical strategies and results. Section 4 discusses implications for firms, in particular for personnel evaluation.

2 Conceptual Framework

2.1 Judging Figure Skating and Figure Skating Judges

In figure skating, a panel of judges and a referee are selected by the ISU (International Skating Union) for each competition from a pool of international judges who are recommended by

national federations. Within approximately 30 seconds after each performance the judges submit two scores–a technical score and an artistic score–that are combined to form the total score. Each score is then displayed on a scoreboard for public viewing.[8]

Judges are monitored and assessed by the ISU. Following each competition, the judges are critiqued by referees, and if a judge is in substantial disagreement with the others on the panel, he or she must be able to defend the deviant score [Yamaguchi *et al.* 1997]. The referee will submit a report that will be the basis of post-competition discussion in the "Event Review Meeting." This referee report is supposed to state any mistakes by the judges and note whether these mistakes have been admitted. The report also includes complaints from other judges or from skaters and coaches. In the meeting, all the judges must respond to every question raised by the referee, skaters, coaches, or other judges. The so-called "acceptable range" of score is determined for each performance, and judges must provide a plausible explanation for any mark outside the range. Those who do not attend the meeting or cannot answer questions are penalized. Since they are unpaid volunteers, the penalty is a written warning or a ban from the next competition [ISU Communication no.1025 1995]. The ISU may informally punish "noisy" judges by not assigning them to major competitions, such as Olympic games and World Figure Skating Championships.[9] The following three types of scoring are considered unsatisfactory [ISU Communication 1999]: (i) systematically deviant scores (e.g. high score for skaters from specific countries), (ii) extraordinary deviation from other judges' scores, and (iii) repeated large deviations from other judges' scores. It seems reasonable to assume that, given the judge-assessment system, it is privately optimal for judges to eschew deviant scoring, and to tend toward agreement.[10]

2.2 Modeling Outlier Aversion Bias

Suppose that there are J experienced judges, $j = 1, 2, \ldots, J$. Without loss of generality, we assume that $J = 2$. Each judge observes the performance of a skater p (Y_{jp}). Skaters are labelled according to their starting order, $p = 1, 2, \ldots, P$. We assume that the starting order is randomly assigned.[11] Before the competition, judges share public information about each skater's quality (Q_p). Performance is observed with error. Specifically, judge j's observation is different from Q_p due to unexpected performance, α_p, by skater p and individual judge-

specific observation error, ϵ_{jp}. Therefore,

$$Y_{jp} = Q_p + \alpha_p + \epsilon_{jp}, \tag{1}$$

where ϵ_{jp} is normally distributed across skater-judge combinations and with mean zero and variance σ^2_ϵ. In other words, a perception error is independently drawn from the identical distribution for each judge-skater pair. Unexpected performance, α_p, is also a random variable and normally distributed across skaters with mean zero and variance σ^2_α. Note that unexpected performance is a common shock to all judges. We assume that α_p and ϵ_{jp} are uncorrelated.

The above specification implies two additional underlying assumptions. First, as already mentioned, judges share common information on the skaters' quality based on their past performances. Second, internationally-qualified judges are homogeneous in the sense that their observations *a priori* suffer the same degree of perception error.

After each skater finishes performing, each judge simultaneously submits a score (S_{jp}), without knowing the other judges' scores. It seems reasonable to assume that judges get utility from a socially acceptable final ranking of the skater and also from scores that are similar to those of the other judges on the panel. The second component of utility is obvious in view of the judge-assessment system. On the other hand, the first motivation leads judges to submit scores that are similar to that of an "average" spectator. Even though amateur spectators might observe more noisy performances than well-trained judges, the average of their observations may as well be $Q_p + \alpha_p$ by the law of large numbers. For individual judge j, the maximum likelihood estimate of the average spectator's observation is his or her own observation (Y_{jp}). The first motivation therefore means that the judges have an incentive to submit scores according to their observations.

Formally, a judge's objective is to balance the trade-off between two factors: (i) minimizing the difference between her score and her own observation and (ii) minimizing the deviation of her score from the other judge's score.[12] We specify utility as dependent on the squared distance between Y_{jp} and S_{jp} to reflect factor (i), and on the squared distance between the judge's score, S_{jp}, and the other judge's score, S_{-jp}, to reflect factor (ii). The simplest

functional form is

$$V_{jp} = -(S_{jp} - Y_{jp})^2 - \lambda_{jp}(S_{jp} - S^e_{-jp})^2, \tag{2}$$

where λ_{jp} is the "price" for deviation from the other judge.[13] The parameter λ_{jp} is specific to each skater-judge combination because it depends not only on a given judge's preferences, but it is also history-dependent. In other words, the degree of agreement with the judge 2 in previous performances changes the current price faced by judge 1. For example, if a judge recently submitted an extreme score, then that judge's price of deviation for a subsequent skater rises. As a result, it is possible to interpret λ_{jp} as the judge's probability of being punished in the end, calculated at the moment of scoring skater p. In this case, it is obvious that it is dependent upon the judge's past deviations. The price for the deviation from her own observation is set to one as a numeraire.

One last point to note is that the λ's are assumed to be known to all judges. This assumption seems reasonable, since international judges are quite homogeneous in preferences, and since judges' scoring of previous skaters is publicly observable after each performance. However, note that the λ's are unknown to the econometrician. Indeed the existence of the λ is what this study attempts to test.

Since the other judge's score is *ex ante* unobserved, judge j forms the conditional expectation of S_{-jp} given available information. Suppose that the judge guesses S_{-jp} as a weighted sum of Y_{-jp} and Q_p. Temporarily assume that the weight is known to the judge as μ_{-jp}, where $\mu_{-jp} \in [0, 1]$.[14] Then the guess is as follows:

$$S_{-jp} = \mu_{-jp} Y_{-jp} + (1 - \mu_{-jp}) Q_p. \tag{3}$$

Using the law of conditional expectation, we have:

$$\begin{aligned} S^e_{-jp} &= E(S_{-jp}|Y_{jp}, \Theta_p) \\ &= \mu_{-jp} E(Y_{-jp}|Y_{jp}, \Theta_p) + (1 - \mu_{-jp}) Q_p \\ &= \mu_{-jp}(\theta Y_{jp} + (1 - \theta) Q_p) + (1 - \mu_{-jp}) Q_p \\ &= \mu_{-jp}\theta Y_{jp} + (1 - \mu_{-jp}\theta) Q_p. \end{aligned}$$

where

$$\theta = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}. \tag{4}$$

Θ_p denotes a set of common knowledge, $\Theta_p = \{Q_p, \alpha_{jp} \sim N(0,\sigma_\alpha^2), \epsilon_{jp} \sim N(0, \sigma_\epsilon^2), \lambda_{jp}\}$. Straightforward calculations yield the optimal score:

$$S_{jp}^* = \frac{1 + \lambda_{jp} \cdot \mu_{-jp} \cdot \theta}{1 + \lambda_{jp}} Y_{jp} + \frac{\lambda_{jp} \cdot (1 - \mu_{-jp} \cdot \theta)}{1 + \lambda_{jp}} Q_p, \tag{5}$$

for $j = 1, 2$. Assuming that the proposed linear scoring strategy is reciprocally rational in equilibrium, we have explicit solutions for the weight on a single judge's observation relative to public information:

$$\mu_{jp} = \frac{1 + \lambda_{-jp} + \lambda_{jp} \cdot \theta}{1 + \lambda_{-jp} + \lambda_{jp} + (1 - \theta^2) \cdot \lambda_{jp} \cdot \lambda_{-jp}}, \tag{6}$$

for $j = 1, 2$. Notice that the optimal weight is known and can be inferred from common knowledge. Furthermore we have the following useful results:

$$\frac{\partial \mu_{jp}}{\partial \lambda_{jp}} < 0, \frac{\partial \mu_{jp}}{\partial \lambda_{-jp}} < 0, \frac{\partial \mu_{jp}}{\partial \theta} > 0. \tag{7}$$

The interpretation is straightforward. If the deviation from a judge's score becomes more costly, the judge tends to put more weight on Q_p. In other words, as a judge's price increases, he or she will become more biased toward the public information. Similarly, if a (second) judge's marginal cost of deviating from S_{jp} increases, the other (first) judge also tends to make her score closer to Q_p. Finally, when errors are relatively more likely to come from judges' misperceptions than a skater's erratic performance, judges are more biased toward public information.

The above comparative statics provide us with intuitive results. Notice that the expected squared deviation of judge j is:

$$\begin{aligned} E[D_{jp}] &= E[S_{jp}^* - S_{-jp}^*]^2 \\ &= E[(\mu_{jp} Y_{jp} + (1 - \mu_{jp}) Q_p) - (\mu_{-jp} Y_{-jp} + (1 - \mu_{-jp}) Q_p)]^2. \end{aligned}$$

Plugging (1) into the above, we have:

$$\begin{aligned} E[D_{jp}] &= E[\mu_{jp}\epsilon_{jp} - \mu_{-jp}\epsilon_{-jp}]^2 \\ &= \mu_{jp}^2 E(\epsilon_{jp})^2 + \mu_{-jp}^2 E(\epsilon_{-jp})^2 \\ &= (\mu_{jp}^2 + \mu_{-jp}^2)\sigma_\epsilon^2 < 2 \cdot \sigma_\epsilon^2. \end{aligned}$$

The expected squared deviation is smaller than $2 \cdot \sigma_\epsilon^2$, which is the expected squared deviation when the judges score independently. Contrary to conventional wisdom, this result implies that a high degree of agreement among evaluators might reflect strategic manipulations due to their aversion to extreme scores. This seems paradoxical, because judges' concerns about the judge-assessment system and its implications for their careers force them to put less weight on their own observation and rely more on public information.[15] The above inequality suggests a surprising result that inter-judge difference is not a good measure of the validity of judgment.

The model also suggests that

$$\frac{\partial E(D_{jp})}{\partial \lambda_{jp}} = 2(\mu_{jp}\frac{\partial \mu_{jp}}{\partial \lambda_{jp}} + \mu_{-jp}\frac{\partial \mu_{-jp}}{\partial \lambda_{jp}})\sigma_\epsilon^2 < 0. \tag{8}$$

$$\frac{\partial E(D_{jp})}{\partial \lambda_{-jp}} = 2(\mu_{jp}\frac{\partial \mu_{jp}}{\partial \lambda_{-jp}} + \mu_{-jp}\frac{\partial \mu_{-jp}}{\partial \lambda_{-jp}})\sigma_\epsilon^2 < 0. \tag{9}$$

$$\frac{\partial E(D_{jp})}{\partial \theta} = 2(\mu_{jp}\frac{\partial \mu_{jp}}{\partial \theta} + \mu_{-jp}\frac{\partial \mu_{-jp}}{\partial \theta})\sigma_\epsilon^2 > 0. \tag{10}$$

The first two predictions are related to outlier aversion bias. Also note that they are empirically testable. First of all we can calculate D_{jp} for each judge-skater combination simply by comparing individual score with the average of the other judges on the same panel. So the problem is whether there is any variation in λ's.

3 Empirical Analysis

3.1 Data and Descriptive Statistics

The data used are scores given by individual judges for figure skating performances in the World Figure Skating Championships in the three seasons, 2001-2003. Each Championship

consists of four events: men, ladies, pairs, and ice dancing. Each event is composed of three programs: preliminary, short, and long. The World Figure Skating Championships requires qualification in the short program, and skaters perform their free-style skating in the qualifying program. For each program, there is a panel of judges composed of one referee and, before 2003, seven to nine judges. The assignment of judges is determined by the ISU, taking into consideration the balance of national representation.

TABLE 1 illustrates the sample structure. All the data are available on the International Skating Union (ISU) official website (`www.isu.org`) or the United States Figure Skating Association (USFSA) website (`www.usfsa.org`). We collected the scoring data on 283 "men" performances, 289 "ladies" performances, and 438 "pairs" and "ice dancing" performances. These numbers amount to 411 judge-program combinations and 9,573 scorings. A judge on average scores about 23 performances in a game. This means that we can follow up a specific judge's judging behavior over about 23 different performances.

As mentioned before, the ISU recently adopted a new judging system that introduced anonymity and random selection of judges. Anonymity prevents the public from specifically identifying the marks awarded by judges. Scores are displayed on the scoreboard in the numerical order. There are, for example, 14 judges in a panel instead of 7 because a computer will randomly select only 7 out of 14 marks for the final ranking. The public cannot identify which marks are selected out of those on the scoreboard. Two results of the reform are notable in our sample. First, the average number of judges in a panel increased from 8.1 to 12.4. It is also now impossible to combine the technical and artistic score of each judge. As a result, when we compare the scores before and after the reform, we must use technical and artistic scores, separately.

Some information on skater quality is available from the so-called crystal reports, such as years of skating experience and rankings in past major competitions. We decided not to use athletic experience as a measure of skater quality because the self-reported years of experience seem to be very noisy. On the other hand, rankings in past major competitions are informative and reliable. In the full sample, those skaters who have been ranked at least once within the top five in the past four years in World Figure Skating Championships ("top five skaters") consist of about 18 percent of total observations. "Top ten skaters" make up

roughly 32 percent.

TABLE 2 presents the means and standard deviations of the average scores by the panel. Some interesting patterns related to subjective performance evaluation are notable. First, artistic score is categorically higher than technical score. Given that artistic scores are presumably more subjective, this implies the presence of leniency bias in figure skating judging. Judges like to look generous at poorly-performing skaters, and they can manipulate artistic scores more easily than technical scores. Second, the standard deviation is consistently larger for technical scores than for artistic scores. Notice that the standard deviation measures the dispersion of average scores across performances because we use panel-average scores, $\overline{S}_p = \frac{1}{J_p}\sum_{j=1}^{J_p} S_{j,p}$. In other words, it represents the extent to which each performance is distinctly scored. Larger standard deviations for technical scores accord with the well-know fact in the literature that there is more significant differentiation between performances when judges rate performers on well-defined specific characteristics [Borman 1982]. In sum, the simple statistics in TABLE 2 already show that subjective evaluations are prone to strategic manipulation.

TABLE 3 shows some interesting patterns. We regress individual scores on various characteristics of judges and performances. First, rankings in past competitions measure skaters' quality quite successfully. *Top Ten* increases the score by 1.3, and *Top Five* does so further by 0.4. The gains come a little bit more from technical scores, which again shows that judges do not differentiate performances in artistic scores. TABLE 3 also shows that there exists nationalistic bias in figure skating judging. Judges favor co-patriotic skaters by about 0.3. It seems quite large compared to the estimate of Zitzewitz [2002], 0.17, and that of Campbell and Galbraith [1996], 0.07. The results also show that artistic scores, the more subjective, are a bit more prone to the bias.

Other findings are also noteworthy. Female judges seem to be more generous than male judges. Men's scores are higher than ladies' and pairs'; and scores in the short program are higher than those in the preliminary round, with those in the free program being highest. Scores also get higher as starting order increases. The last finding reflects that skaters are seeded in the free program.

3.2 Identification I: Dynamic Panel Data Estimation

To reiterate, the purpose of this study is to examine whether judges submit scores in a strategic fashion, particularly avoiding outlying scores. The model implies that a direct test would be to calculate whether λ is zero or not. While this calculation does not seem implementable, there are indirect ways of testing for outlier aversion. Notice that the prediction in equation (8) is true only when there exists non-zero outlier aversion. It is possible to test equation (8) if there is any exogenous variation in λ. In this subsection, we exploit dynamic variation in λ_{jp} over multiple performances within one program, λ_{j1}, λ_{j2}, ..., λ_{jP}. Recall that a judge's "price" for deviation from others may well change over the course of performances; it will increase if the judge has already submitted outlying scores for previous skaters. It will be constant or decrease as the judge has been well in concurrence with other judges. The basic empirical specification is therefore dynamic:

$$D_{j,p} = \alpha_1 D_{j,p-1} + \cdots + \alpha_L D_{j,p-L} + \beta \overline{D}_j^{p-L-1} + \gamma Q_{j,p} + \delta p + \tilde{\lambda}_j + \nu_{j,p}, \tag{11}$$

where

$$D_{j,p} = (S_{j,p} - \bar{S}_{-j,p})^2, \tag{12}$$

and

$$\overline{D}_j^{p-L-1} = \frac{1}{p-L-1}\left(\sum_{k=1}^{p-L-1} D_{j,k}\right). \tag{13}$$

$\overline{D}_j^{p-L-1}$ measures the average squared deviation of individual judge j's score from the other judges' average score up to the $(p-L-1)$-th previous skater, and $Q_{j,p}$ is a vector of the skater's quality and a constant term. It includes dummy variables which are equal to one if the skater was at least once ranked within top five or top ten in World Figure Skating Championships for the past four years and zero otherwise. Starting order, p, is included in case there exists any related systematic effect.[16]

For identification, we have to assume that the deviations before the L-th previous performance have only an "average" marginal effect, β, while the previous deviations up to the L-th previous performance have idiosyncratic effects on current scoring. This assumption seems

reasonable in the sense that judges have a "fresher memory" about recent performances. We will estimate the simplest case of equation (11) for $L = 1$.[17]

The $\nu_{j,p}$'s are assumed to have finite moments and, in particular, $E(\nu_{j,p}) = E(\nu_{j,p}\nu_{j,q}) = 0$ for $p \neq q$. In other words, we assume the absence of serial correlation but not necessarily independence over starting order. The autocorrelation structure is testable [Arellano and Bond 1991]. It is also assumed that the initial conditions of the dependent variable, $D_{j,1}$, are uncorrelated with the subsequent disturbances $\nu_{j,p}$ for $p = 2, \ldots, P$. The initial conditions are predetermined. However the correlation between $D_{j,1}$ and $\tilde{\lambda}_j$ is left unrestricted.

The parameter $\tilde{\lambda}_j$ is an unobserved judge-specific skater-invariant effect that allows for heterogeneity in the means of the series of $D_{j,p}$ across judges. We treat the judge-specific effects as being stochastic. There are two interpretations for $\tilde{\lambda}_j$.[18] First, the effect may represent the individual judge's risk aversion that affects his or her aversion to outlying scores. Judges might be heterogeneous in their career concerns. Those judges who like to pursue their career as judges are more likely to be conservative and would be more averse to outlying scores.[19] Note that λ_{jp} is unobservable to econometricians and is specified as random.

Second, the judge-specific effect may represent an idiosyncratic benchmark point of scoring. Figure-skating judges have only to rank skaters relatively. As a result, absolute values of scores do not matter much.[20] For example, suppose that a judge mistakenly scores the first skater higher in absolute terms than do all the other judges. If the judge tries to adjust his or her initial mistake in absolute terms and rank the subsequent skaters accordingly (with the same inflation), then that judge's deviations will be larger than those of the other judges for all skaters in the program. In this case, the following deviations reflect only the initial deviation and have nothing to do with outlier aversion. We must allow for the individual-specific intercepts of deviations in order to test for outlier aversion.

Because of the presence of $\tilde{\lambda}_j$, lagged dependent variables included in the right-hand side in equation (11) are necessarily endogenous unless the distribution of $\tilde{\lambda}_j$ is degenerate. However, with the assumptions on the error terms, it is possible to estimate the coefficients in equation (11) consistently in two steps: (i) eliminate $\tilde{\lambda}_j$ by the first-differencing transformation and (ii) use the values of the dependent variable lagged by two skaters or more as instrumental

variables (in the case that $L = 1$). This is the "Arellano-Bond GMM estimator" or the "GMM-DIF estimator" [Arellano and Bond 1991].[21] Specifically, we will estimate the following first-differenced equation:

$$\Delta D_{j,p} = \alpha \Delta D_{j,p-1} + \beta \Delta \overline{D}_j^{p-2} + \gamma \Delta Q_{j,p} + \delta + \Delta \nu_{j,p}, \tag{14}$$

where Δ represents the first-differencing transformation. We assume that $Q_{j,p}$ is strictly exogenous. Notice that $\overline{D}_j^{p-2}$ is predetermined, so its lagged values can be used as additional instruments. The key identifying assumption is that the lagged level $D_{j,p-k}$ will be uncorrelated with $\triangle \nu_{j,p}$ for $k \geq 2$, together with the assumption of initial conditions.

Before progressing further, one might suspect that the squared deviation, although convenient for analysis, is really what judges are concerned about. Fortunately, it is possible to conduct a direct test of whether $D_{j,p}$ is meaningful for the judge-assessment system and judges' career concerns. We estimate a simple probit that regresses $R_{j,2002}$ on $D_{j,p,2001}$, where $R_{j,2002}$ is a dummy variable that is one if the judge j is re-selected for the 2002 Championships conditional on the fact that the judge is selected in 2001, and zero otherwise.[22] Judges' nationality is controlled for to take into account that each national federation has its own unique procedure of recommending judges to the ISU.[23] After controlling for country-specific effects, we find that an increase in the average degree of squared deviation per performance (about 0.13) significantly reduces the probability of reselection by about 1 percent (p-value = 0.03). Thus, if a judge continued to deviate by the average for 20 skaters (the average number of skaters that a typical judge is supposed to score), then the probability of reappointment will decrease by more than 20 percent. It is obvious that volunteer judges are honored to be selected for major international competitions, like the World Figure Skating Championships. This implies that the squared deviation should be one of the important statistics in the judge-assessment system that judges are concerned about.

The coefficients, α and β, are of major interest. We expect that their signs will be negative, since previous deviations would lead judges to avoid further deviations. The coefficients for the skater's quality are presumably negative in part because top skaters' performances are stable and less extraordinary (larger θ) and in part because judges' evaluations are more conservative for these top skaters (larger λ). Greater attention is paid to those top performances

by the media, spectators, and therefore the ISU. Furthermore, the ISU judge-monitoring system explicitly puts more weight on serious bias or error for highly-ranked skaters [ISU Communication no.1197].

We estimate equation (11) using ordinary least squares (OLS), a within-group estimator (WG), and the Arellano-Bond dynamic panel data (GMM) model and juxtapose the estimates for comparison. Even though this model can be consistently estimated only by GMM, the comparison with these potentially inconsistent estimates may be very useful. The asymptotic results and Monte-Carlo studies have shown that the OLS estimator is biased upward and the WG estimator is biased downward if $|\alpha| \leq 1$ [Blundell, Bond, and Windmeijer 2000].[24] Therefore, if the empirical model is correctly specified and there is no finite sample bias, any consistent estimate must lie between the corresponding OLS and WG estimates. Whether this pattern (sometimes called the "Sevestre-Trognon inequality") is observed or not is a simple and valid test for specification and finite sample biases [Bond 2002, Sevestre and Trognon 1997].

TABLE 4 presents the estimation results for the sample for the 2001 and 2002 seasons.[25] We also run the same regressions separately for sub-samples, singles (men and ladies) and pairs (pairs and ice dancing), the results of which are presented in TABLE 5 and 6, respectively.

Before discussing the estimates, let us examine some of the specification issues mentioned above. First, we find across-the-board that the OLS estimates for the lagged dependent variable appear to be larger than the corresponding GMM estimates, while the WG estimates appear to be smaller. For the full sample, when skaters' quality is controlled, the GMM estimate is -0.0375 – between the OLS estimate (0.0516) and the WG estimate (-0.0807).

The relationship between the estimates confirms the Sevestre-Trognon inequality. The bias in the WG estimates is small relative to that of the OLS estimates. It is a well known fact that the asymptotic bias of the WG estimate is inversely related to the length of time period. By the within-group transformation, the lagged dependent variable becomes $D_{j,p-1} - \frac{1}{P}(D_{j,1} + \cdots + D_{j,p} + \cdots + D_{j,P})$, and the error term becomes $\nu_{j,p} - \frac{1}{P}(\nu_{j,1} + \cdots + \nu_{j,p-1} + \cdots + \nu_{j,P})$. These two are obviously correlated, above all because $D_{j,p-1}$ and $\frac{1}{P}\nu_{j,p-1}$ are correlated, $\frac{1}{P}D_{j,p}$ and $\nu_{j,p}$ are correlated, and so on. For sufficiently large P, the correlations will be negligible. The

"time" period in this paper is quite long, about 23 skaters in a typical program. Indeed we find the size of the bias in the WG estimates is relatively small. The validity of the instruments is strongly supported by the Sargan test of over-identifying restrictions (the p-value is higher than 0.99 for every specification).

Finally, the assumption of no serial correlation of ν_{jp} cannot be rejected. The last two rows in the tables present the Arellano-Bond test statistics for autocorrelation. We find that there is significant negative first-order serial correlation in the first-differenced residuals, while there is no second-order correlation. It is consistent with the assumption that the error term in level is serially uncorrelated. The AR(1) structure is accepted at a p-value lower than 0.01, and the AR(2) structure is rejected across the board at a p-value higher than 0.50.

The GMM estimates imply that the deviation of a judge's vote for the previous skater significantly decreases the deviation for the current skater. This result is consistent with equation (8). Suppose that the judge's score is deviated from the average of the others by the extent of 0.45.[26] The estimates imply that the deviation pressures judges to be biased by about 0.09 point ($=\sqrt{\alpha \times 0.45^2}=\sqrt{0.038 \times 0.45^2}$) toward the average for the current player. Similarly, the outlier aversion bias to the average deviation amounts to 0.11 ($=\sqrt{0.056 \times 0.45^2}$) for the singles competition and 0.11 ($=\sqrt{0.06 \times 0.45^2}$) for the pairs and ice dancing competitions as seen in TABLE 5 and 6.[27]

Based on the idea put forth by Campbell and Galbraith [1996], the size of the bias can be explained in the following way: imagine a judge who has difficulty in deciding between two neighboring scores, separated by 0.1.[28] Suppose that there exists a critical value for that judge's previous deviation, beyond which he or she will choose the score closer to the average for the current situation. If the previous deviation is less than the critical value, the judge then randomizes her score between the two neighboring scores. Such a judge shows a bias of 0.05 in response to the critical value. The estimated size of the bias is economically significant. It is interesting to compare these estimates with those of nationalistic bias: Zitzewitz [2002] finds that nationalistic bias is on average 0.17, and Campbell and Galbraith [1996] find nationalistic bias of 0.07 by nonparametric estimation.

The marginal effect of the average squared deviation up to the $(p-L-1)$-th previous performance is larger than that of the one-time deviation for the immediately preceding

performance. In other words, $|\beta| > |\alpha|$. For the full sample, when skaters' quality is controlled, β (-0.1658) is almost five times as large as α (-0.0375) in absolute terms. This result seems reasonable, because β picks up a kind of cumulative effect of α.[29] For example, the magnitude of β for the full sample implies that if one judge deviated from the other judges' average by 0.45, then the current score is likely to be closer to the average by 0.18 ($=\sqrt{0.1658 \times 0.45^2}$). This again confirms the existence of outlier aversion.

The estimates for Q_p are also consistent with the model's predictions. Judges are more in agreement for top skaters. If a skater was at least once ranked within top ten in the past four Championships, then the squared deviation decreases by about 0.07 to 0.1. As mentioned before, it is in part because top skaters are less erratic, and also in part because the price of the deviation for the judge is higher when evaluation top-ranked skaters due to the judge-assessment system. Both explanations are consistent with the conceptual model.

TABLE 7 presents the results when the effects of positive and negative deviations are separated out to test for symmetry. Overall, the results are very similar. Once a judge has submitted outlying scores, she is more likely to converge toward the group. Also, judges are more in agreement for top-ranked skaters. Interestingly, the GMM estimates suggest that scoring should be a little more responsive to positive deviations than negative deviations, even though the effects are not statistically different. This indicates that judges are more averse to positive extreme scores. In other words, they may be more afraid of scoring too high. The finding makes sense when one considers that positive bias is usually considered as favoritism, which is a more sensitive issue in this sport.

In TABLE 8 we re-estimate the GMM model on technical and artistic scores separately. As found in TABLE 2 and 3 artistic scores are more prone to bias because they are more subjective than technical scores. Thus we expect that outlier aversion bias should be larger in artistic scores. The result shows that artistic scores are indeed more responsive to previous deviations. Again, it implies that judges can manipulate artistic score more easily than technical score. They like to avoid outlying judgment in total score, and they do by adjusting artistic score.

3.3 Identification II: Interim Judging System

In this section, we exploit the quasi-natural experiment of the judging system reform in 2002 to examine outlier aversion. In 2002 the ISU adopted a new system called the Interim Judging System in which judges' names are concealed on the scoreboard from outside observers, including judges themselves. Also a judge's score is randomly selected for the final ranking. The new system was implemented in the World Figure Skating Championships in 2003.

The change in the judging system provides another opportunity to test the existence of outlier aversion bias, since one might expect that judges would be less pressured to agree under the new system. The ISU itself states "anonymity reduces the risk of judges coming under external pressure." The "external pressure" referred to by the ISU is mainly assumed to be nationalistic favoritism. However, it is important to note that anonymity relieves judges of the stress exerted by another source of external pressure, the media and fans, which is not negligible at all in this sport. Olympic gold medalist Carol Jenkins said "people watching at home will be ready in their mind to do their own judging. It's the one sport where the spectators judge the judges." Indeed historic scoring scandals have been initially provoked by the media and fans rather than by the ISU itself. We expect that the introduction of anonymity, though cannot remove completely, significantly weakens judges' incentives for outlier aversion. As a result, a meaningful test for the existence of outlier aversion bias is whether judges' scores became more dispersed after the introduction of anonymity. The launch of the new system provides unique opportunity to exploit a natural experiment in the area of personnel policy.

TABLE 9 shows the simple mean comparisons of deviations before and after the system changed. For robustness, we use three measures of score dispersion for the same skater:

$$\xi_p^1 = \frac{1}{J_p - 1} \sum_{j=1}^{J_p} (S_{j,p} - \overline{S}_p)^2,$$

$$\xi_p^2 = \frac{2}{J_p(J_p - 1)} \sum_{i=1}^{J_p} \sum_{j=1}^{J_p} |S_{i,p} - S_{j,p}|,$$

$$\xi_p^3 = S_p^{max} - S_p^{min}.$$

The first measure (ξ_p^1) is the consistently estimated standard deviation of the sample; the

second measure (ξ_p^2) is the average absolute deviation; the last measure (ξ_p^3) is the range between the maximum and minimum score. The number of judges in a panel (J_p) is subscripted by p, because it varies over skaters. Note that the measures of dispersion are standardized with respect to number of judges except ξ_p^3.

In TABLE 9 all the measures increased under the new system in 2003. For the men's program, the standard deviation of technical scores increased from 0.16 to 0.18, and the range increased from 0.47 to 0.60. For the ladies' program, the standard deviation of artistic scores increased from 0.18 to 0.20, and the range increased from 0.52 to 0.65. Most of these changes are statistically significant at reasonable levels. Thus we conclude that the new system seems to reduce judges' outlier aversion bias.

The simple comparison of means is intuitive, but one can object that it does not control for other variables. Above all, the increases in dispersion might reflect aggravated nationalistic bias and an increase in corrupt scoring after the reform. Indeed, the new system has been harshly criticized in that it could allow judges to manipulate their scores more easily without accountability.[30] To meet this objection, we regress the amount of dispersion on several control variables, including a measure of nationalistic bias (an indicator of whether the skater and at least one judge on the panel are from same country (B_p)). The equation to be estimated is:

$$\xi_p = b + \beta_1 B_p + \beta_2 A_p + \beta_3 B_p A_p + Q_p \gamma + X_p \delta + u_p, \tag{15}$$

where ξ_p is one of the three dispersion measures.[31] A_p is the indicator of anonymity (one for the new system and zero for the old system)[32]; X_p is a vector of indicators for events and programs; Q_p is a vector of measures of skaters' quality.

TABLE 10 shows the results. The estimates of β's are of primary interest, and all should be positive. Indeed, we find across the board that scores are more dispersed after the introduction of anonymity in the judging system, even after controlling for the nationalistic bias. Furthermore, the standard deviations significantly increase by about 0.01, the absolute deviations increase by about 0.1, and the range increases by about 0.3. The magnitude ranges from 13 to 36 percent of one standard deviation of each measure.[33]

Let us call the panel with at least one judge from the same country as the skater the

"co-patriotic" panel, and the panel without any such judge the "neutral" panel. Scores of co-patriotic panels are slightly more convergent, although not statistically significant. On the other hand, we find strong evidence of nationalistic bias. Both maximum and minimum scores are higher for co-patriotic panels. Furthermore the nationalistic bias is aggravated under the new system. As a result, the votes of co-patriotic panels are significantly more dispersed under the Interim Judging System.

Other results are consistent with the predictions of the model. First, scores are significantly more convergent for top skaters. For all three measures, the dispersion of scores shrinks by half a standard deviation. Second, scores in more advanced programs are also less divergent.

4 Measuring Information Loss

So far we have found that there is an incentive for judges to agree on their scoring based on the particular judge-assessment system in the sport of figure skating. Individual judges' strategic scoring to avoid outlying judgments distorts the distribution of scores across judges, slanted toward a single reference point based on public information and compressed around that point. Given that scores are the weighted sums of private signals and public information, the distortion of the score distribution implies some loss of private information contained in individual observation. An interesting question here is to assess how much information on the margin is discarded due to outlier aversion.

We employ two approaches to approximate the size of the information loss.[34] Both are based on the idea that we can measure the loss by comparing scores with and without outlier aversion. However notice that the idea is not fully-implementable since we can only observe scores tainted with outlier aversion bias. For the first measure, we first construct score deviations that should be made by a hypothetical judge who does not have outlier aversion. The gap between these deviations and those actual deviations represents the size of the bias toward public information or the amount of private information discarded for convergence. Note that the hypothetical deviations are supposed to be larger. Specifically, imagine a panel of judges who do not have outlier aversion at all. If a judge on this panel deviated from the

other judges, this would be not because of his or her strategic manipulation, but because of pure individual-specific observation error. As a result, the hypothetical judge's deviation is also a random variable. Suppose that the judge's squared deviation is on average 0.0256, and assume there are 20 skaters in a program.[35] Since the judge's current scoring is independent of previous deviations (no strategic response), the expected value of the average squared deviation for 20 performances is simply 0.0256.

What would have happened to the expected average squared deviation if the judge did have outlier aversion of the same degree we found in the sample? Assume that the squared deviation for each of the first two performances is 0.0256. In other words, judges are initially same as the hypothetical judge, and there is no outlier aversion. However, thereafter, judges' scores would not be independent of their previous deviations, even though the initial deviations are randomly assigned. The data-generating process, specified by equation (11), shows the progress of the following deviations. The imputation based on the GMM estimates and initial values yields 0.0214 as the average for 20 performances.[36] The ratio of the expected average squared deviations is about $0.84 = \frac{0.0214}{0.0256}$. This implies that approximately 16 percent of information in terms of squared deviation would be lost due to outlier aversion.

Another measure of the information loss can be obtained by comparing score dispersion before and after the judging-system reform. The underlying idea is that score dispersion among different judges reflects the relative inclusion of private observation. If they all submitted consensus opinions based on public information, the score distribution would be degenerate. The new judging system is supposed to weaken outlier aversion, so we expect larger dispersion after the reform. The increase in score dispersion indicates that some information was discarded by compression due to outlier aversion before the reform.[37] A simple measure to compare score dispersion before and after the reform is the ratio of variances:

$$L = \frac{\mathrm{Var}(S_{j,p,t<2003})}{\mathrm{Var}(S_{j,p,t=2003})}. \tag{16}$$

It is straightforward to estimate L from previous estimates. In TABLE 10 we know that standard deviation increases by 0.0131 after the reform. Since the average standard deviation before the reform was about 0.165 in TABLE 9, an estimate of L is therefore about $0.86 = (\frac{0.1650}{0.1650+0.0131})^2$. This implies that the information loss amounts to approximately 14 percent

in terms of variance, very similar to the other estimate of 16 percent.

5 Implications

The purpose of this paper is to illustrate that subjective evaluators are sensitive to the incentive structure they work in and that they are likely to be biased depending on how they are monitored and assessed. We focus on a specific kind of bias, outlier aversion bias in subjective evaluations in presence of multiple evaluators. The case of figure skating judging clearly shows that there is a bias toward agreement, because the degree of agreement among judges is used as a measure of the reliability of the evaluations and to assess individual judges themselves.

These findings have interesting implications for group decision-making in business and organizational contexts. When deciding to implement subjective evaluations, it is important to take into account the system used to assess the evaluators. Employing multiple evaluators cannot prevent individualistic bias and error if they interact with each other in a strategic way. They will cooperate and manipulate their decisions as long as there exists a mutually-beneficial externality in the incentive structure. It is as important to prevent collusive behavior as to devise way to aggregate different preferences and minimize idiosyncratic errors in subjective evaluation.

The results also imply that agreement among evaluators is not always desirable. Firms often utilize subjective evaluation in group decision-making process. Unfortunately, as the findings have demonstrated, these processes are subject to outlier aversion bias because of the incentives faced by members of the decision-making group. For example, evaluators may not want to upset their bosses or hold up a time-sensitive decision. When firms gather groups for input and decision-making, they may believe that those processes result in an accurate compilation of beliefs from those who are involved and informed. It is, however, likely that the outcomes of those meetings are biased toward consensus, do not accurately reflect the true opinions of the participants, and may harm firms because misinformation brings about misjudgment, especially when the pending issue is very important and decision-makers feel pressured to reach a concrete, unified decision. When agreement is externally induced, this

often leads to a loss of valuable private information that individual evaluators may have had access to but that the others do not. Multiple evaluators aggregate to make more accurate judgments because individual observational errors are cancelled out by integrating different opinions. However, it should be noted that valuable private information is weighted less when the diversity of opinion is averaged out. Objections to the consensus by credible informants should be encouraged rather than offered disincentives.

References

[1] Arellano, Manuel and Bond, Stephen. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equation." *Review of Economic Studies*, 1991, 58, pp. 277-297.

[2] Bassett, Gilbert W. and Persky, Joseph. "Rating Skating." *Journal of the American Statistical Association*, 1994, 89, pp. 1075-1079.

[3] Bernheim, B. Douglas. "A Theory of Conformity." *Journal of Political Economy*, 1994, 102(5), pp. 1075–1079.

[4] Blundell, Richard, Bond, Stephen and Windmeijer, Frank. "Estimation in Dynamic Panel Data Models: Improving on the Performance of the Standard GMM Estimators." in Badi Baltagi (ed.), *Advances in Econometrics, Volume 15: Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, Amsterdam: JAI Elsevier Science, 2000.

[5] Bond, Stephen. "Dynamic Panel Data Models: A Guide to Micro Data Methods and Practice." *Portugese Economic Journal*, 1, 2002, 141-162.

[6] Campbell, Bryan and Galbraith, John W. "Non-parametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments." *Statistician*, 1996, 45, 521-526.

[7] Garicano, Luis, Palacios, Ignacio and Prendergast, Canice. "Favoritism under Social Pressure." *NBER Working Paper*, no.8376, July 2001.

[8] Ginsburgh, Victor A. and van Ours, Jan C. "Expert Opinion and Compensation: Evidence from a Musical Competition." *American Economic Review*, March 2003, 93(1), 289-296.

[9] Janis, Irving L. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, Boston: Houghton Miffin Company, Second Edition, 1983.

[10] Lamont, Owen A. "Macroeconomic Forecasts and Microeconomic Forecasters." *Journal of Economic Behavior and Organization*, July 2002, 48(3), 265-280.

[11] Milkovich, George T. and Wigdor, Alexandra K. eds. *Pay for Performance: Evaluating Performance Appraisal and Merit Pay*, Washington D.C.: National Academy Press, 1991.

[12] Murphy, Kevin R. and Cleveland, Jeanette N. *Performance Appraisal: An Organizational Perspective*. Massachusetts: Allyn and Bacon, 1991.

[13] Nickell, Stephen J. "Biases in Dynamic Models with Fixed Effects." *Econometrica*, 1981, 49(6), 1417-1426.

[14] Prendergast, Canice. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 1999, 37(1), 7-63.

[15] Prendergast, Canice. "A Theory of "Yes Men."" *American Economic Review*, September 1993, 83(4), 757-770.

[16] Saal, Frank, Downey, Ronald and Lahey, Mary Anne. "Rating the Ratings: Assessing the Quality of Rating Data." *Psychological Bulletin*, September 1980, 88(2), 413-428.

[17] Saxonhouse, Gary R. "Estimated Parameters as Dependent Variables." *American Economic Review*, March 1976, 66(1), 178-183.

[18] Sevestre, Patrick and Trognon, Alain. "Dynamic Linear Models." in *The Econometrics of Panel Data,* Boston and London: Kluwer Academic, 1996.

[19] Topel, Robert and Prendergast, Canice. "Favoritism in Organizations." *Journal of Political Economy*, October 1996, 104(5), 958-978.

[20] Weekley, Jeff A. and Gier, Joseph A. "Ceilings in the Reliability and Validity of Performance Ratings: The Case of Expert Raters." *Acedemy of Management Journal*, 1989, 32(1), 213-222.

[21] Yamaguchi, Kristi, Ness, Christy and Meacham, Jody. *Figure Skating for Dummies*. IDG Books Worldwide Inc.: Foster City, CA., 1997.

[22] Zitzewitz, Eric. "Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making." *Working Paper*, Graduate Business School, Stanford University, October 2002.

Notes

[1]For general discussion about subjective performance evaluation, see Prendergast [1999], Prendergast and Topel [1996], Baker, Gibbons, and Murphy [1994], and Prendergast [1993] among others.

[2]The bias means that evaluators intentionally do not submit their best estimate or opinion. For example, there are two types of evaluator bias often noted in the literature, "centrality bias" and "leniency bias" [Prendergast 1999]. The first occurs when subjective evaluators do not differentiate accurately between agents, and the latter occurs when evaluators overstate the performance of poor performers. Both types of bias may arise due to the incentive system; taking employee appraisals as an example, first, the supervisor wants to maintain a positive relationship with the employee who is being judged (leading to centrality bias) and, second, also wants to encourage poorly-performing workers through positive reinforcement (leading to leniency bias).

[3]There are many cases where the appraisal is based on multiple sources. Firms gather and integrate information about workers from their coworkers and supervisors. Also, school principals compile their own observations of a student with those from other teachers.

[4]Note that it is different from centrality bias. Outlier aversion bias arises when multiple evaluators tend to submit similar judgment, while centrality bias occurs when an evaluator submits similar judgment for multiple agents or performers.

[5]Janis [1983] documents historical moments such as the Cuban missile crisis and Korean war where conforming to group norms within the president's inner circle, and thereby ignoring minor opinions led to disastrous and irrational consequences. He argues that pressure to conform within small cohesive groups of decision-makers is an important source of faulty decision-making and collective misjudgment.

[6]There are many examples or anecdotes in real life that evaluators strategically pretend to agree, even though they really don't think so. One example is employee performance review. Consider a firm that hires a consultant from outside the firm to assess workers. If the consultant realizes that the firm will be assessing her work by comparing it with other internal reports, then she is likely to to hide whatever private information she may have discovered and, instead, will try to mimic inside analysts' opinion [Prendergast and Topel 1999]. Another example can be found in macroeconomic forecasting. Lamont [2002] empirically finds that forecasters tend to gravitate toward consensus estimates. In particular, the tendency is stronger among younger forecasters, which indicates that forecasters' career concerns are affecting their predictions. It is, however, only partially subjective because forecasts are at least *ex post* verifiable by a third party.

[7]Since the scandal, various reforms of the judging system have been under consideration. One proposal is to specify detailed standards to reduce subjectivity. This approach has been criticized by the commentators who argue that it will change figure skating into a "jumping contest."

[8]This subsection explains the judging system before the 2002 reform. As will be explained, there are some

substantial differences before and after the reform. For example, it becomes impossible to identify a specific judge's score on the scoreboard by the principle of anonymity.

[9]This will be formally tested on our sample.

[10]Another obvious reason why judges are reluctant to submit extreme scores is that they are trained not to do so. Under the U.S. system, candidates must show experience as "trial" judges and then "test" judges to finally become "competition" judges. For promotion, 75 percent of the "trial" judge's rankings must correspond to the regular judges. To be promoted from from "test" to "competition" judges, 90 percent accordance is required.

[11]There are usually three stages in figure skating competition; the preliminary round, short program, and long program. For the long program, the starting order is not randomly determined, but the skaters are seeded after the short program. According to their ranking up to the short program, each of them is assigned to a group. The best group will skate at the end of the long program. However, the order within each group is again determined by a random draw.

[12]The model does not consider favoritism for specific skaters.

[13]Here we assume that the prices of positive deviation and that of negative deviation are equal. Otherwise we may allow for λ_{jp}^{+} and λ_{jp}^{-}, separately. The distinction is empirically testable.

[14]We are looking for μ_{jp}, $j = 1, 2$, an optimal guess in rational expectation equilibrium. My model is an application of Cournot duopoly.

[15]The main objective of training judges and measuring the degree of agreement among them is to reduce their perception error, σ_{ϵ}^{2}. However the effect might be overestimated without taking into account the fact that the type of training and appraisal system would aggravate judges' outlier aversion and thereby increase the degree of agreement.

[16]Starting order is randomly assigned to each skater except in the case of long programs (or free skating programs), in which the order is determined by rankings from previous programs.

[17]In fact increasing L does not make any significant difference in results.

[18]We assume that a judge may have different judge-specific effects in different programs. For example a judge, Mr. Fairmind, is treated as two different judges when he judges for men's short program and for men's free program.

[19]Unfortunately, there is no available information about judges' characteristics but nationality.

[20]Then, one might expect that what judges really care about should be their ranking, not scores. However, the ISU explicitly mentions they investigate the marks awarded.

[21]We refer to this as the "GMM estimator."

[22]The selection of judges in the 2003 Championships is not considered because the total number of judges

selected in that year increased due to the judging-system reform.

[23]However, the results do not change when country-specific effects are not controlled for.

[24]The WG estimator eliminates $\tilde{\lambda}_p$ by transforming the original observations in to deviations from individual means. However this transformation induces a non-negligible correlation between the transformed lagged dependent variable and the transformed error term.

[25]Remember that the data on individual scorings are not available for 2003 because of anonymity under the new judging system.

[26]It is the average of the distance between median and extrema.

[27]The best way to evaluate the size of the bias is to measure the gap between actual observation and the corresponding submitted score. However, it is impossible because we do not have any information on judges' actual observations.

[28]The unit of score is 0.1.

[29]The WG estimates for β are categorically downward biased. This is consistent with Nickell [1981].

[30]For related criticisms refer to http://skatefair.visionsnet.com.

[31]Robust standard errors are calculated since the dependent variable is the estimated parameter [Saxonhouse 1976].

[32]It is, therefore, simply the yearly dummy variable with one for 2003 and zero for 2001 and 2002. Separating 2001 and 2002 does not change the following results.

[33]In each case, "one standard deviation" is: 0.08 for the standard deviation, 0.098 for the absolute deviation, and 0.26 for the range.

[34]We ignore the quality issue of information on the margin.

[35]The number is chosen because it is the average squared deviation for the sample. However we can choose any different number.

[36]We generate the squared deviations for skater $p = 3, 4, \cdots, 20$, according to the estimated equation, $D_{jp} = \hat{\alpha}_1 D_{j,p-1} + \hat{\beta}\overline{D}_j^{p-2} + \tilde{\lambda}_j$. Therefore $E(D_{jp}) = \hat{\alpha}_1 D_{j,p-1} + \hat{\beta}\overline{D}_j^{p-2} + E(\tilde{\lambda}_j)$. We assume that the first two deviations, $E(D_{j1})$ and $E(D_{j2})$, are 0.0256 because the estimation does not have any systematic behavioral implication for the initial deviations.

[37]Note that the second measure is likely to underestimate the information loss because the reform does not completely remove outlier aversion. The unbiased measure should be to compare score dispersion with and without outlier aversion.

Table 1: **Number of observations: skater-judge combinations**

Event	Program	2001	2002	2003
Men	Qualifying	301	266	400
	Short	270	270	420
	Long	216	216	216
Ladies	Qualifying	329	280	399
	Short	270	261	420
	Long	216	207	336
Pairs	Qualifying	-	-	-
	Short	216	180	294
	Long	171	180	280
Ice Dance	Qualifying	490	392	261
	Short	270	252	406
	Long	216	216	336
Total		2,965	2,720	3,888

Table 2: **Panel-average scores**[1]

	2001		2002		2003	
	Technical	Artistic	Technical	Artistic	Technical	Artistic
Full Sample	4.705	4.889	4.668	4.856	4.695	4.873
	(.704)	(.629)	(.715)	(.641)	(.716)	(.656)
Men	4.841	5.003	4.868	4.995	4.773	4.972
	(.594)	(.517)	(.639)	(.562)	(.665)	(.598)
Ladies	4.713	4.885	4.561	4.821	4.655	4.845
	(.655)	(.584)	(.729)	(.608)	(.709)	(.615)
Pairs	4.621	4.826	4.610	4.792	4.665	4.820
	(.779)	(.704)	(.730)	(.697)	(.760)	(.724)
Qualifying	4.593	4.733	4.534	4.662	4.571	4.678
	(.738)	(.683)	(.748)	(.725)	(.765)	(.728)
Short	4.657	4.965	4.620	4.938	4.659	4.948
	(.712)	(.580)	(.741)	(.565)	(.743)	(.622)
Long	4.963	5.069	4.921	5.047	4.887	5.018
	(.559)	(.521)	(.556)	(.508)	(.576)	(.546)

[1] Panel-average score is $\overline{S}_p = \frac{1}{J_p} \sum_{j=1}^{J_p} S_{j,p}$. Standard deviations are displayed in parentheses.

Table 3: **Preliminary look at scores in level**[1]

	Total	Technical	Artistic
Top Five	.4226	.2259	.1967
	(.0310)	(.0177)	(.0147)
Top Ten	1.292	.6613	.6308
	(.0296)	(.0166)	(.0144)
Co-patriotic Judge	.2723	.1247	.1476
	(.0585)	(.0322)	(.0279)
Female Judge	.0838	.0460	.0378
	(.0255)	(.0143)	(.0121)
Ladies	-.4837	-.2818	-.2019
	(.0338)	(.0192)	(.0159)
Pairs	-.4948	-.2675	-.2274
	(.0308)	(.0169)	(.0148)
Short Program	.0830	-.0542	.1371
	(.0331)	(.0184)	(.0153)
Free Program	.4858	.2473	.2384
	(.0313)	(.0168)	(.0153)
Starting Order	.0397	.0212	.0184
	(.0020)	(.0011)	(.0010)
Constant	8.695	4.277	4.418
	(.0400)	(.0218)	(.0192)
$R^2 =$	.4979	.4700	.5012

[1] Number of observations is 5,685 scores from 2001 and 2002. Robust standard errors are displayed in parentheses.

Table 4: **Dynamic panel data estimation: full sample**[1]

	OLS		WG		GMM	
	(1)	(2)	(3)	(4)	(5)	(6)
$D_{j,p-1}$	.0671	.0516	-.0703	-.0807	-.0378	-.0375
	(.0184)	(.0180)	(.0147)	(.0145)	(.0183)	(.0180)
$\overline{D}_j^{p-2}$	.3024	.3258	-.6787	-.6700	-.1698	-.1658
	(.0464)	(.0466)	(.0574)	(.0564)	(.0821)	(.0812)
Top Five		-.0087		-.0162		-.0138
		(.0073)		(.0132)		(.0812)
Top Ten		-.0955		-.0991		-.0997
		(.0079)		(.0110)		(.0133)
Starting Order		-.0017				
		(.0006)				
Constant	.0897	.1471	.2465	.2851	.0035	.0012
	(.0066)	(.0107)	(.0095)	(.0098)	(.0009)	(.0009)
Observations	4,782	4,782	4,782	4,782	4,540	4,540
Number of judges	242	242	242	242	235	235
Sargan test					232.89	231.79
AR(1) test					-38.25	-37.63
AR(2) test					.44	.68

[1] Robust standard errors are displayed in parentheses.

Table 5: **Men and ladies**[1]

	OLS		WG		GMM	
	(1)	(2)	(3)	(4)	(5)	(6)
$D_{j,p-1}$	.0536	.0358	-.0647	-.0721	-.0559	-.0576
	(.0201)	(.0200)	(.0208)	(.0207)	.0258)	(.0257)
$\overline{D}_j^{p-2}$	.3065	.3519	-.6170	-.6123	-.1406	-.1655
	(.0508)	(.0513)	(.0850)	(.0844)	(.1285)	(.1282)
Top Five		-.0199		-.0300		-.0197
		(.0165)		(.0161)		(.0181)
Top Ten		-.0369		-.0429		-.0256
		(.0141)		(.0138)		(.0164)
Starting Order		-.0031				
		(.0008)				
Constant	.0894	.1452	.2248	.2471	-.0031	-.0023
	(.0084)	(.0121)	(.0127)	(.0132)	(.0010)	(.0010)
Observations	2,490	2,490	2,490	2,490	2,362	2,362
Number of judges	128	128	128	128	121	121
Sargan test					117.10	116.84
AR(1) test					-28.02	-28.86
AR(2) test					.17	.14

[1] Robust standard errors are displayed in parentheses.

Table 6: **Pairs and ice dancing**[1]

	OLS		WG		GMM	
	(1)	(2)	(3)	(4)	(5)	(6)
$D_{j,p-1}$	.0754	.0614	-.0740	-.0858	-.0598	-.0559
	(.0260)	(.0252)	(.0210)	(.0205)	(.0243)	(.0238)
$\overline{D}_j^{p-2}$	.2982	.3069	-.7163	-.7103	-.3352	-.2941
	(.0619)	(.0604)	(.0798)	(.0776)	(.1035)	(.1015)
Top Five		-.0281		-.0346		-.0335
		(.0085)		(.0200)		(.0224)
Top Ten		-.1307		-.1286		-.1485
		(.0118)		(.0158)		(.0184)
Starting Order		-.0006				
		(.0010)				
Constant	.0904	.1539	.2655	.3222	-.0040	-.0002
	(.0098)	(.0183)	(.0145)	(.0150)	(.0014)	(.0014)
Observations	2,292	2,292	2,292	2,292	2,178	2,178
Number of judges	114	114	114	114	107	107
Sargan test					110.99	113.80
AR(1) test					-26.19	-25.34
AR(2) test					.14	.42

[1] Robust standard errors are displayed in parentheses.

Table 7: **Positive and negative deviation**[1]

	OLS	WG	GMM
$D_{j,p-1}$	.0515	-.0809	-.0497
	(.0180)	(.0145)	(.0179)
$\overline{D}_j^{p-2}(+)$	.3615	-.6185	-.2869
	(.0727)	(.0858)	(.1769)
$\overline{D}_j^{p-2}(-)$	.3060	-.7053	-.1820
	(.0494)	(.0717)	(.1125)
Top Five	-.0086	-.0162	-.0120
	(.0073)	(.0132)	(.0152)
Top Ten	-.0955	-.0991	-.1006
	(.0080)	(.0110)	(.0132)
Starting Order	-.0017		
	(.0006)		
Constant	.1466	.2844	-.0013
	(.0107)	(.0099)	(.0009)
Observations	4,782	4,782	4,540
Number of judges	242	242	235
Sargan test			818.47
AR(1) test			-37.48
AR(2) test			.56

[1] Robust standard errors are displayed in parentheses.

Table 8: **Technical and artistic scores**[1]

	GMM	
	Technical	Artistic
$D_{j,p-1}$	-.0149	-.0694
	(.0190)	(.0178)
$\overline{D}_j^{p-2}$	-.1748	-.2206
	(.0958)	(.0769)
Top Five	.0197	-.0038
	(.0118)	(.0111)
Top Ten	-.0435	-.0266
	(.0136)	(.0128)
Constant	-.0007	-.0014
	(.0003)	(.0003)
Observations	4,540	4,540
Number of judges	235	235
Sargan test	$p > .99$	$p = .97$
AR(1) test	$p < .01$	$p < .01$
AR(2) test	$p = .77$	$p = .88$

[1] Robust standard errors are displayed in parentheses.

Table 9: **Interim judging system: before-after analysis**[1]

	Technical			Artistic		
Men	Before	After	△	Before	After	△
Standard deviation	.1589	.1825	.0236	.1629	.1698	.0068
	[.0657]	[.0753]	(.0087)	[.0663]	[.0698]	(.0084)
Absolute deviation	.1793	.2056	.0263	.1836	.1913	.0077
	[.0765]	[.0889]	(.0102)	[.0793]	[.0799]	(.0100)
Range	.4651	.5989	.1339	.4672	.5553	.0881
	[.2007]	[.2674]	(.0284)	[.2015]	[.2308]	(.0267)
Ladies						
Standard deviation	.1766	.2069	.0303	.1785	.1986	.0201
	[.0860]	[.0723]	(.0102)	[.0781]	[.0806]	(.0099)
Absolute deviation	.1995	.2338	.0342	.2022	.2238	.0216
	[.1010]	[.0841]	(.0120)	[.0925]	[.0938]	(.0116)
Range	.5192	.6740	.1548	.5192	.6479	.1287
	[.2644]	[.2556]	(.0327)	[.2318]	[.2660]	(.0304)
Pairs						
Standard deviation	.1748	.1770	.0021	.1638	.1837	.0199
	[.0948]	[.0818]	(.0097)	[.0951]	[.0923]	(.0101)
Absolute deviation	.1969	.1999	.0030	.1835	.2051	.0216
	[.1124]	[.0961]	(.0115)	[.1122]	[.1063]	(.0118)
Range	.5095	.5805	.0710	.4721	.6008	.1287
	[.2771]	[.2672]	(.0292)	[.2817]	[.3050]	(.0307)

[1] Standard deviations are in brackets. Robust standard errors are displayed in parentheses.

Table 10: **Interim judging system: regression analysis**[1]

	Standard	Absolute	Range		
	Deviation	Deviation		Max	Min
Interim System	.0131	.0128	.0932	-.0064	-.0996
	(.0048)	(.0056)	(.0155)	(.0327)	(.0381)
Co-patriot	-.0038	-.0053	-.0203	.0891	.1094
	(.0042)	(.0048)	(.0125)	(.0255)	(.0312)
Interim System	.0095	.0153	.0494	.1062	.0568
× Compatriot	(.0069)	(.0080)	(.0227)	(.0463)	(.0544)
Top 5	-.0371	-.0435	-.1260	.2471	.3731
	(.0047)	(.0055)	(.0150)	(.0217)	(.0309)
Top 10	-.0422	-.0491	-.1306	.6362	.7668
	(.0041)	(.0047)	(.0129)	(.0210)	(.0276)
Ladies	.0228	.0263	.0733	-.1586	-.2319
	(.0041)	(.0048)	(.0131)	(.0283)	(.0332)
Pairs	.0128	.0142	.0358	-.2524	-.2882
	(.0040)	(.0047)	(.0125)	(.0252)	(.0305)
Qualify	.0465	.0594	.1007	-.1641	-.2648
	(.0039)	(.0046)	(.0123)	(.0250)	(.0304)
Short	.0247	.0292	.0758	-.0948	-.1706
	(.0036)	(.0041)	(.0118)	(.0246)	(.0292)
Artistic	-.0035	-.0045	-.0153	.1734	.1887
	(.0033)	(.0038)	(.0103)	(.0206)	(.0248)
Constant	.1544	.1721	.4728	4.8904	4.4176
	(.0044)	(.0051)	(.0142)	(.0301)	(.0359)
R^2 =	.2287	.2377	.2385	.4394	.4650
Observations			2,020		

[1] Robust standard errors are displayed in parentheses.