

Extreme Sports Evaluation: Evidence from Judging Figure Skating

Jungmin Lee*

First version: April 20, 2003

This version: June 1, 2004

Abstract

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 “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 quasi-natural experiment shows that the dispersion of scores across judges depends upon the type of judge-assessment system and its implication for outlier aversion. One major conclusion of this paper, which is at variance with the industrial psychology and personnel management literature, is that agreement may not be a good criterion for the validity of an evaluation system.

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 a panel of judges. In these cases, objective measures are either impractical or distorted, making subjective measures the only

*Email: junglee@eco.utexas.edu. Department of Economics, University of Texas, Austin, TX 78712. I would like to thank Dan Hamermesh, Gerald Oettinger, Max Stinchcombe, Steve Trejo, Justin Wolfers, Eric Zitzewitz, and seminar participants at the University of Texas.

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 or distorted by the evaluators who are pursuing their own goals other than unbiased review. 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 allows preventing individualistic bias, for example nepotism, now that the organization can detect extraordinary 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. This paper calls this tendency of gravity "outlier aversion bias."⁴ They suppress the

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

²The 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).

³There 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. **Multi-source performance appraisals?**

⁴Note 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.

desire to question alternatives in favor of agreement seeking. They would not want to be branded as “nonconformists” (Bernheim 1994). In particular, when evaluators are appraised through comparison with their peer evaluators, there will likely be 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.⁵ 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. It is found that 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, we 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 judgment process is almost entirely subjective, and 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.⁶

Our 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

⁵There 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 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.

⁶Since 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.”

distribution will likewise change in a systematic fashion. In this paper, we exploit two kinds of variation in the judges' incentive: (i) endogenous variation of an individual judge's averseness to outlying judgments 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 in 2002. We expect to find a systematic change in voting behavior, which represents a similar change in the judges' incentive structure.

We find evidence that 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. This paper overall confirms that subjective evaluations are sensitive to the 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 businesses.

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.⁷

Judges are monitored and assessed by the ISU. Following each competition, the judges are

⁷This 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.

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 given (Yamaguchi *et al.* 1997). The referee will submit a report that will be the basis of discussion in a post-competition meeting called the “Event Review Meeting.” The referee report is supposed to state any mistakes by the judges and note whether these mistakes have been admitted. The referee report also includes complaints from other judges or from skaters and coaches. In this meeting, 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 warnings or a ban from the next competitions (ISU Communication no.1025 1995). More important than formal penalties, the ISU may indirectly punish problematic judges by not assigning them to major competitions such as Olympic games and World Figure Skating Championships.⁸ It therefore seems reasonable to assume that given the judge-assessment system, it is privately optimal for judges to eschew deviant scorings, and to tend toward agreement.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. This paper focuses on the last two types of bias. In this system, there is significant incentive for outlier aversion.

2.2 Modeling Outlier Aversion Bias

Suppose that there are J experienced judges, $j = 1, 2, \dots, J$. 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, $p = 1, 2, \dots, P$. We assume that the starting order is randomly assigned.¹⁰ Before the competition, judges share public information about each

⁸This will be formally tested on our sample.

⁹Another 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 experience “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 of the accordance is required.

¹⁰There 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

skater’s quality (Q_p). Performance is observed with errors. 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,

$$Y_{jp} = Q_p + \alpha_p + \epsilon_{jp}, \quad (1)$$

where ϵ_{jp} is normally distributed across skater-judge combinations and with mean zero and variance σ_ϵ^2 . 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_α^2 . 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, judges as internationally-qualified ones are homogenous in the sense that their observations *a priori* suffer the same degree of perception errors.

After each skater finishes performance, judges simultaneously submit 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 number. 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 the 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.¹¹ 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

$$V_{jp} = -(S_{jp} - Y_{jp})^2 - \lambda_{jp}(S_{jp} - S_{-jp}^e)^2, \quad (2)$$

again determined by a random draw.

¹¹The model does not consider favoritism for specific skaters.

where λ_{jp} is the “price” for deviation from the other judge.¹² 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 becomes larger. 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 λ ’s are assumed to be known to all judges. This assumption seems reasonable since international judges are quite homogenous in preferences, and since judges’ scoring of previous skaters is publicly observable after each performance. However, note that λ ’s are unknown to econometricians. Indeed the existence of λ is what this paper 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 $\mu_{-jp} \in [0, 1]$.¹³ Then the guess is as follows:

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

Using the law of conditional expectation, we have:

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

where

$$\theta = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}. \quad (4)$$

And Θ_p denotes a set of common knowledge, $\Theta_p = \{Q_p, \alpha_{jp} \sim N(0, \sigma_\alpha^2), \epsilon_{jp} \sim N(0, \sigma_\epsilon^2)\}$,

¹²Here 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.

¹³We are looking for μ_{jp} , $j = 1, 2$, an optimal guess in rational expectation equilibrium. My model is an application of Cournot duopoly.

λ_{jp} }. Straightforward calculations yield the optimal score:

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

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:

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

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:

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

If the deviation from the second judge's score becomes more costly, the first 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 the second judge's marginal cost of deviating from S_{jp} increases, this judge also tends to make her score closer to Q_p . Finally, when errors are relatively more likely to come from judges' misperception 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:

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

This inequality suggests the surprising result that inter-judge difference is not a good measure of the validity of judgment. The expected squared deviation is smaller than $2 \cdot \sigma_\epsilon^2$, which is the expected squared deviation when the judges independently score. 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.¹⁴

Furthermore the model suggests the following:

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

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

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

In particular, the first two predictions are related to outlier aversion bias. For empirical purposes, we can calculate D_{jp} for each judge-skater combination. As a result, if there is any variation in λ , then it is possible to test predictions for (8) and (9). As is explained in detail in the next section, we exploit two kinds of variation in carrying out these tests: (i) endogenous changes in λ over multiple skaters for an individual judge and (ii) exogenous and across-the-board changes in λ due to the judging system reform in 2002.

3 Empirical Analysis

3.1 Data and Descriptive Statistics

The data used are scores for figure skating performances given by individual judges in the World Figure Skating Championships in the three seasons between 2001 and 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.

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.usfssa.org).

¹⁴The 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.

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. TABLE 1 illustrates the sample structure.

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. And 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. Second, it is no longer possible to identify which judge submitted each score. 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, such as years of skating experience and rankings in past major competitions. 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 of 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 might imply the presence of leniency bias. In other words, subjective evaluators are more generous. Second, the standard deviation is consistently larger for technical score than for artistic score. Notice that the standard deviation measures the dispersion of average scores across performances. In other words, it represents the extent to which each performance is distinctly scored. Larger standard deviations for technical scores are in accordance with the well-know fact in the literature that there is significant differentiation between performances when judges rate performers on well-defined specific characteristics (Borman 1982). Simple descriptive statistics already show that subjective evaluations are prone to strategic manipulation.

3.2 Identification I: Dynamic Panel Data Estimation

To reiterate, the purpose of this paper is to examine whether judges submit scores in a strategic fashion, particularly eschewing outlying scores. The model from the previous section 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. Notice that 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, $\lambda_{j1}, \lambda_{j2}, \dots, \lambda_{jP}$. Recall that a judge’s “price” for deviation from others may as 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 as follows:

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

where

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

and

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

\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.¹⁵

For identification, we have to assume that the deviations before the L -th previous performance have only “average” marginal effect, β , while the previous deviations up to the L -th previous performance have idiosyncratic effects on the 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 $L = 1$.¹⁶

¹⁵Starting 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.

¹⁶In fact increasing L does not make any significant difference in results.

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 lack 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 $p = 2, \dots, P$. 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 basically two interpretations for the judge-specific effect, $\tilde{\lambda}_j$.¹⁷ First, it is related to the “time-constant” part of λ_{jp} in the theoretical model that is supposed to be constant over starting order within a program. The judge-specific effect may represent the individual judge’s risk averseness that affects his or her averseness to outlying scores. Judges might be heterogeneous in their concerns about judge career. Those judges who like to pursue their career as judges are more likely to be conservative. They would be more averse to outlying scores.¹⁸ Note that λ_{jp} is unobservable to econometricians and is specified as random.

Second, we need to include the existence of judge-specific effects because figure-skating judges have only to rank skaters relatively. In this sport, absolute values of scores do not matter much.¹⁹ For example, suppose that a judge mistakenly scores the first skater higher than do all the other judges. If the judge tries to be consistent – because of relative ranking – 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 dispersion of initial scores and have nothing to do with outlier aversion. Therefore, judge-specific effects must be controlled for 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

¹⁷We 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.

¹⁸Unfortunately, there is no available information about judges’ characteristics but nationality.

¹⁹Then, 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.

(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 called the “Arellano-Bond GMM estimator” or the “GMM-DIF estimator” (Arellano and Bond 1991, Blundell and Bond 1998).²⁰ Specifically, we will estimate the following first-differenced equation:

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

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

Before progressing further, one might suspect that the squared deviation, though 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 model 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.²¹ Judges’ nationality is controlled for to take into account that each national federation has its own unique procedure of recommending judges to the ISU.²² After controlling for country-specific effects, we find that 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 should have decreased 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 at least 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

²⁰We refer to this as the “GMM estimator.”

²¹The 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.

²²However, the results do not change when country-specific effects are not controlled for.

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 square (OLS), a within-group estimator (WG), and the Arellano-Bond dynamic panel data (GMM) model, of which the estimates are juxtaposed for comparison. Even though this model can be consistently estimated only by GMM, the comparison with these potentially inconsistent estimates is very useful from the econometric point of view. 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).²³ 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 “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 3 presents the estimation results for the sample for the 2001 and 2002 seasons.²⁴ 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 4.6, 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 $D_{j,p-1} - \frac{1}{P}(D_{j,1} + \dots + D_{j,p} + \dots + D_{j,P})$, and the error term becomes $\nu_{j,p} - \frac{1}{P}(\nu_{j,1} + \dots + \nu_{j,p-1} + \dots + \nu_{j,P})$. 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

²³The 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.

²⁴Remember that the data on individual scorings are not available for 2003 because of anonymity under the new judging system.

“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). The estimates imply that the deviation by the average difference between median and maximum (or minimum) score (about 0.45) 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 6.²⁵ 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 0.1.²⁶ 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. And if the previous deviation is less than the critical value, the judge then randomizes her score between the two neighboring scores. Such a judge show 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

²⁵The 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.

²⁶The unit of score is 0.1.

reasonable, because β picks up a kind of cumulative effect of α .²⁷ 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 4 presents the results when the effects of positive and negative deviations are separated out. 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 current scoring should be more responsive to positive deviations than negative deviations, even though the effects are not statistically different. This result indicates that judges are more averse to positive extreme scores. In other words, they may be more afraid of scoring too high. This makes sense when one considers that positive bias is usually considered as favoritism, which is a more sensitive issue in this sport.

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

²⁷The WG estimates for β are categorically downward biased. This is consistent with Nickell (1981).

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

$$\begin{aligned}\xi_p^1 &= \frac{1}{J_p - 1} \sum_{j=1}^{J_p} (S_{j,p} - \bar{S}_p)^2, \\ \xi_p^2 &= \frac{2}{J_p(J_p - 1)} \sum_{i=1}^{J_p} \sum_{j=1}^{J_p} |S_{i,p} - S_{j,p}|, \\ \xi_p^3 &= S_p^{max} - S_p^{min}.\end{aligned}$$

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 7 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 significance levels. Thus we conclude that the new system seems to reduce judges' outlier aversion bias.

The simple mean comparison 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 scorings 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.²⁸ To meet this objection, we regress the amount of dispersion on several

²⁸For related criticisms refer to <http://skatefair.visionsnet.com>.

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:

$$\xi_p = b + \beta_1 B_p + \beta_2 A_p + \beta_3 B_p A_p + Q_p \gamma + X_p \delta + u_p, \quad (15)$$

where ξ_p is one of the three dispersion measures.²⁹ A_p is the indicator of anonymity (one for the new system and zero for the old system)³⁰; X_p is a vector of indicators for events and programs; Q_p is a vector of measures of skaters’ quality.

TABLE 8 shows the results. The estimates of β ’s are of primary interest. We expect that all the β ’s 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.³¹

Let us call the panel with at least one judge from the same country as the skater the “compatriotic” panel, and the panel without any such judge the “neutral” panel. Scores of compatriotic panels are slightly more converged, although not statistically significant. On the other hand, we find strong evidence of nationalistic bias. Both maximum and minimum scores are higher for compatriotic panels. Furthermore the nationalistic bias is aggravated under the new system. As a result, the votes of compatriotic panels are significantly more dispersed under the Interim Judging System.

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

4 Information Loss on the Margin

We find 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

²⁹Robust standard errors are calculated since the dependent variable is the estimated parameter (Saxonhouse 1976).

³⁰It 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.

³¹In each case, “one standard deviation” is: 0.08 for the standard deviation, 0.098 for the absolute deviation, and 0.26 for the range.

to avoid outlying judgment makes the distribution of scores across judges compressed and slanted toward a single reference point based on public information. Given that scores are the weighted sum of observation and public information, outlier aversion bias leads to some loss of private information contained in individual observation. An interesting question is to assess how much information is discarded due to outlier aversion.

We employ two approaches to approximate the size of the information loss.³² Both are based on the idea that we can measure the loss by comparing scores with and without outlier aversion bias. More specifically, the first one is based on the hypothesis that a judge should submit his or her actual observations if there were no incentive for outlier aversion. The gap between scores submitted by the hypothetical judge without outlier aversion and those submitted by actual judges represents the size of the bias toward public information. In other words, the gap represents the amount of information discarded for convergence. For calculation, imagine a judge who does not have outlier aversion at all. If this judge deviated from the other judges, this would be not because of his or her strategic manipulation, but because of pure individual-specific observation error. The hypothetical judge's deviation is nothing but a random variable. Suppose that the judge's squared deviation is on average 0.0256.³³ And assume there are 20 skaters. Since the judge's current scoring is independent of previous deviations (no strategic response), the expected value of 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. However the data-generating process, specified by equation (11) with the GMM estimates, yields 0.0214 as the average for 20 performances.³⁴ It is obvious that, on average, the deviation decreases because the judge avoids future outlying scores after the previous deviations. The ratio of 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) is lost due to the existence of outlier aversion.

Another crude measure of the information loss can be obtained by comparing score dispersion before and after the change in the judging system. The underlying idea is that the

³²We ignore the quality issue of information on the margin.

³³The number is chosen because it is the average squared deviation for the sample.

³⁴We generate the squared deviations for skater $p = 3, 4, \dots, 20$, according to the estimated equation, $D_{jp} = \hat{\alpha}_1 D_{j,p-1} + \hat{\beta} \bar{D}_j^{p-2} + \tilde{\lambda}_j$. Therefore $E(D_{jp}) = \hat{\alpha}_1 D_{j,p-1} + \hat{\beta} \bar{D}_j^{p-2} + E(\tilde{\lambda}_j)$. 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.

diversity of scores among judges reflects their relative inclusion of private observation. Recall that, in our model, the parameter of outlier aversion ultimately represents the relative weight on public information. If they all submitted consensus opinion based on public information, there should be no diversity at all. As seen before, the new judging system is supposed to weaken outlier aversion among judges, so we expect to see more diversity after the reform. Judges are relieved from conformity pressure, and so they now tell what they really observe. The increased dispersion of scores indicates that so much information on the margin was discarded due to outlier aversion before the reform.³⁵ A simple measure to compare diversity in scores before and after the reform is the ratio of variances (score dispersion):

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

It is straightforward to estimate L from TABLE 7 and 8. In TABLE 8 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 7, an estimate of L is therefore about $0.8583 = \left(\frac{0.1650}{0.1650+0.0131}\right)^2$. This implies that the information loss is approximately 14 percent in terms of variance, which is very similar to the first estimate.

5 Implications

The purpose of this paper is to illustrate that subjective evaluators are sensitive to the incentive structure they work in. In particular, subjective evaluations are dependent on how the evaluators are appraised and how their evaluations are perceived. The case of figure skating judging clearly shows that there is a bias toward agreement, because the degree of agreement among judges is used, one, as a measure of the reliability of the evaluations and, two, to assess individual judges themselves.

From these findings, it is possible to derive some interesting implications for group decision-making in business and organizational contexts. First, when deciding to implement subjective evaluations, it is important to take into account the system used to assess the evaluators. In the case of multiple evaluators, they will cooperate and manipulate their decisions as long as there exists mutually-beneficial externality in the incentive structure. It is as important to prevent collusive behaviors as to devise how to aggregate different preferences.

³⁵Note 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.

Second, my results imply that agreement among evaluators is not always desirable. Firms often utilize subjective evaluation in group decision-making process. Unfortunately, as my 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; however, it is 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 and mismanagement.³⁶ In those cases, 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 valued, not discouraged.

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.” *Working Paper*, 2000.

³⁶Janis (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 decision-making errors.

- [5] Bond, Stephen. "Dynamic Panel Data Models: A Guide to Micro Data Methods and Practice." *Working Paper*, April 2002.
- [6] Campbell, Bryan and Galbraith, John W. "Non-parametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments." *Statistician*, 1996, 45, pp. 521-526.
- [7] Garicano, Luis, Palacios, Ignacio and Prendergast, Canice. "Favoritism under Social Pressure." *Working Paper*, 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), pp. 289-296.
- [9] Janis, Irving L. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, Boston: Houghton Mifflin Company, Second Edition, 1983.
- [10] Lamont, Owen A. "Macroeconomic Forecasts and Microeconomic Forecasters." *Journal of Economic Behavior and Organization*, July 2002, 48(3), pp. 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), pp. 1417-1426.
- [14] Prendergast, Canice. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 1999, 37(1), pp. 7-63.
- [15] Prendergast, Canice. "A Theory of "Yes Men."" *American Economic Review*, September 1993, 83(4), pp. 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), pp. 413-428.
- [17] Saxonhouse, Gary R. "Estimated Parameters as Dependent Variables." *American Economic Review*, March 1976, 66(1), pp. 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), pp. 958-978.
- [20] Weekley, Jeff A. and Gier, Joseph A. "Ceilings in the Reliability and Validity of Performance Ratings: The Case of Expert Raters." *Academy of Management Journal*, 1989, 32(1), pp. 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.

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

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

Table 3: **Dynamic Panel Data Estimation: Full Sample**¹

	OLS		WG		GMM	
	(1)	(2)	(3)	(4)	(5)	(6)
$D_{j,p-1}$.0671 (.0184)	.0516 (.0180)	-.0703 (.0147)	-.0807 (.0145)	-.0378 (.0183)	-.0375 (.0180)
\overline{D}_j^{p-2}	.3024 (.0464)	.3258 (.0466)	-.6787 (.0574)	-.6700 (.0564)	-.1698 (.0821)	-.1658 (.0812)
Top Five		-.0087 (.0073)		-.0162 (.0132)		-.0138 (.0812)
Top Ten		-.0955 (.0079)		-.0991 (.0110)		-.0997 (.0133)
Starting Order		-.0017 (.0006)				
Constant	.0897 (.0066)	.1471 (.0107)	.2465 (.0095)	.2851 (.0098)	.0035 (.0009)	.0012 (.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 4: **Dynamic Panel Data Estimation:
Full Sample**¹

	OLS	WG	GMM
$D_{j,p-1}$.0515 (.0180)	-.0809 (.0145)	-.0497 (.0179)
$\overline{D}_j^{p-2}(+)$.3615 (.0727)	-.6185 (.0858)	-.2869 (.1769)
$\overline{D}_j^{p-2}(-)$.3060 (.0494)	-.7053 (.0717)	-.1820 (.1125)
Top Five	-.0086 (.0073)	-.0162 (.0132)	-.0120 (.0152)
Top Ten	-.0955 (.0080)	-.0991 (.0110)	-.1006 (.0132)
Starting Order	-.0017 (.0006)		
Constant	.1466 (.0107)	.2844 (.0099)	-.0013 (.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 5: **Dynamic Panel Data Estimation: Singles**¹

	OLS		WG		GMM	
	(1)	(2)	(3)	(4)	(5)	(6)
$D_{j,p-1}$.0536 (.0201)	.0358 (.0200)	-.0647 (.0208)	-.0721 (.0207)	-.0559 (.0258)	-.0576 (.0257)
\overline{D}_j^{p-2}	.3065 (.0508)	.3519 (.0513)	-.6170 (.0850)	-.6123 (.0844)	-.1406 (.1285)	-.1655 (.1282)
Top Five		-.0199 (.0165)		-.0300 (.0161)		-.0197 (.0181)
Top Ten		-.0369 (.0141)		-.0429 (.0138)		-.0256 (.0164)
Starting Order		-.0031 (.0008)				
Constant	.0894 (.0084)	.1452 (.0121)	.2248 (.0127)	.2471 (.0132)	-.0031 (.0010)	-.0023 (.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: **Dynamic Panel Data Estimation: Pairs**¹

	OLS		WG		GMM	
	(1)	(2)	(3)	(4)	(5)	(6)
$D_{j,p-1}$.0754 (.0260)	.0614 (.0252)	-.0740 (.0210)	-.0858 (.0205)	-.0598 (.0243)	-.0559 (.0238)
\overline{D}_j^{p-2}	.2982 (.0619)	.3069 (.0604)	-.7163 (.0798)	-.7103 (.0776)	-.3352 (.1035)	-.2941 (.1015)
Top Five		-.0281 (.0085)		-.0346 (.0200)		-.0335 (.0224)
Top Ten		-.1307 (.0118)		-.1286 (.0158)		-.1485 (.0184)
Starting Order		-.0006 (.0010)				
Constant	.0904 (.0098)	.1539 (.0183)	.2655 (.0145)	.3222 (.0150)	-.0040 (.0014)	-.0002 (.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: **Interim Judging System: Before-After Analysis**¹

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

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

Table 8: **Interim Judging System: Regression Analysis**¹

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

¹ Robust standard errors are displayed in parentheses.

Table 9: **Determinants of Scores in Level**¹

	Total	Technical	Artistic	Total	Technical	Artistic
Top Five	.4226 (.0310)	.2259 (.0177)	.1967 (.0147)	.3967 (.0434)	.2094 (.0239)	.1873 (.0207)
Top Ten	1.292 (.0296)	.6613 (.0166)	.6308 (.0144)	1.300 (.0369)	.6671 (.0203)	.6328 (.0176)
Compatriotic Judge	.2723 (.0585)	.1247 (.0322)	.1476 (.0279)	.2600 (.0622)	.1162 (.0342)	.1438 (.0297)
Female Judge	.0838 (.0255)	.0460 (.0143)	.0378 (.0121)			
Ladies	-.4837 (.0338)	-.2818 (.0192)	-.2019 (.0159)			
Pairs	-.4948 (.0308)	-.2675 (.0169)	-.2274 (.0148)			
Short Program	.0830 (.0331)	-.0542 (.0184)	.1371 (.0153)			
Free Program	.4858 (.0313)	.2473 (.0168)	.2384 (.0153)			
Starting Order	.0397 (.0020)	.0212 (.0011)	.0184 (.0010)	.0437 (.0019)	.0235 (.0142)	.0202 (.0009)
Constant	8.695 (.0400)	4.277 (.0218)	4.418 (.0192)	8.495 (.0258)	4.123 (.0142)	4.371 (.0123)
Judge-Specific Effect	No	No	No	Yes	Yes	Yes
$R^2 =$.4979	.4700	.5012	.4467	.4094	.4549

¹ Number of observations is 5,685 scores. Robust standard errors are displayed in parentheses.