Outlier Aversion in Subjective Evaluation

Evidence from World Figure Skating Championships

AQ1

The quality of subjective performance evaluation is dependent on the incentive structures evaluators face. Figure skating competitions provide a unique opportunity to study subjective evaluation. Using scoring data from World Figure Skating Championships between 2001 and 2003, I test for the existence of "outlier aversion," in which subjective evaluators avoid submitting outlying judgments. I find that judges manipulate scores to achieve a targeted level of agreement with the other judges. Agreement may not be a good criterion for the validity of an evaluation system, consistent with the recent applied psychology and management literature.

Keywords: *subjective performance evaluation; outlier aversion; figure skating*

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 (Prendergast, 1999; Topel & Prendergast, 1996). However, it is well known that there are chronic problems with subjective performance appraisals. First of all these evaluations cannot be verified by any other than evaluators themselves. It is therefore impossible to figure out the underlying processes by which evaluators reach their judgment. Subjective measures can be possibly manipulated by the evaluators who are pursuing goals other than unbiased reviews. Accurate evaluation is often a relatively minor concern of appraisers relative to their own rent seeking. While researchers in personnel psychology have studied rater training and rating format for the reliability of performance appraisals for decades, their focus has recently expanded to cover rater motivation (Murphy & De Schon, 2000; Murphy & Cleveland, 1995; Ilgen, Barnes-Farrell, & McKellin, 1993).

JOURNAL OF SPORTS ECONOMICS, Vol. X No. X, Month 2007 1–19
DOI: 10.1177/1527002507299203

From the economics point of view, it is important to see how the quality of subjective evaluation depends on the incentive system caused by the organization. In particular, from the mechanism design perspective, the organization needs to devise a system that prevents evaluators from manipulating their judgment in an arbitrary way. A simple, and perhaps the most practical, way of checking subjectivity is to employ multiple evaluators (e.g., hire external experts or consultants to supplement internal reviewers) and to compile their opinions. There are two advantages with multiple evaluations. First, they can reduce or prevent individualistic bias, such as nepotic favoritism, because the evaluators know that the organization can detect unusual evaluation by comparing different evaluations. Second, aggregating multiple appraisals can average out individuals' idiosyncratic measurement errors. When different raters independently provide similar ratings for the same performance, it is traditionally accepted as a form of consensual validity or convergent validity in the personnel psychology literature (Viswesvaran, Ones, & Schmidt, 1996).

This paper will show however that evaluators distort their judgment under the multiple appraisal system. The idea is most closely related to that of Prendergast (1993), who, in the context of the principal-agent model, 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, which is based on his or her own observation of the opinions of subordinate evaluators. Along this line of thought, I examine whether subjective evaluators have an incentive to distort their assessments toward a general consensus when they are ex post appraised through comparison with their peer evaluators. I call this tendency toward consensus "outlier aversion."[1]

I use individual judges' scoring data from World Figure Skating Championships between 2001 and 2003. Figure skating is an excellent sport for testing outlier aversion because its judging process is almost entirely subjective.[2] Because of frequent scandals and controversies about scoring, figure skating judges are closely monitored and assessed by their supervising organization. This provides a unique opportunity to test theories about subjective evaluation. The empirical part of this paper consists of two tests. First, I examine whether judges attempt to avoid submitting outlying scores particularly when they have already submitted some extreme ones. Specifically, I measure each judge's deviation from the other judges' scores on the same panel and estimate the autoregressive effects of previous deviations on current scoring using the dynamic panel data model. Second, I exploit a discrete change in the organization and system of evaluation caused by a major Olympic scandal in 2002 involving the gold medal award for pairs skating. The new system introduced anonymity and random selection of scores, both of which reduced the probability of judges being punished on outlying scores. Exploiting exogenous variation in judges' incentive structure due to this natural experiment, I test whether score distribution across judges became more dispersed under the new system.

2. FIGURE SKATING JUDGES

In figure skating, a panel of judges and a referee are selected by the International Skating Union (ISU) for each competition from a pool of international judges who are recommended by national federations. Judges are monitored and assessed by the ISU. For 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. Following each competition, judges are critiqued by their panel referee, 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 (International Skating Union, 1999). As will be shown later, the ISU does punish noisy judges by not assigning them to major competitions, such as Olympic Games and World Figure Skating Championships. The following three types of scoring are considered unsatisfactory: (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. Obviously the purpose of the judge-assessment system is to ensure the objectivity of judgments by minimizing any bias of scores. However the system also provides external and unintended incentives for judges to distort their ratings toward artificial agreement. This study examines whether judges really respond to the incentive system that they face.

3. DATA AND DESCRIPTIVE STATISTICS

The data used are scores given by all the judges in the World Figure Skating Championships in the three seasons from 2001 through 2003. Each championship consists of four events: men, ladies, pairs, and ice dancing. And each event is composed of three programs: preliminary, short, and long program. The World Figure Skating Championships requires qualification in the short program, and skaters perform their freestyle 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. AQ3

All the data are publicly available on the ISU official website (www.isu.org) or the United States Figure Skating Association (USFSA) website (www.usfsa.org). I 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 for a program in an event.

The ISU recently adopted a new (interim) judging system that introduced anonymity and random selection of judges (International Skating Union, 2002). 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, all of them submit their marks and a computer randomly selects only 9 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 the sample. First, the average number of judges in a panel increased from 8.1 to 12.4. Second, it becomes impossible to combine the technical and artistic score of each judge.

Some information on skater quality is available from the so-called crystal reports, including years of skating experience and rankings in past major competitions. I decided not to use athletic experience as a measure of skater quality because the self-reported years of experience seem to be inaccurate. 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 (or ten) in the past four years in World Figure Skating Championships consist of about 18 percent of total observations.

Table 1 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. This might reflect the fact that judges are given more precise guidelines for scoring technical elements. Alternatively, given that artistic scores are presumably more subjective, this implies the presence of leniency bias in figure skating judging. Judges like to look generous in regards to 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. It represents the extent to which each performance is distinctly scored. Larger standard deviations for technical scores accord with the well-known 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 suggest that figure skating scores are prone to strategic manipulation.

To check the data quality, I regress individual scores on various characteristics of judges and performances. Table 2 shows the results. First, rankings in past competitions measure skaters' quality quite successfully. If a skater has been

TABLE 1: 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
	(0.704)	(0.629)	(0.715)	(0.641)	(0.716)	(0.656)
Men	4.841	5.003	4.868	4.995	4.773	4.972
	(0.594)	(0.517)	(0.639)	(0.562)	(0.665)	(0.598)
Ladies	4.713	4.885	4.561	4.821	4.655	4.845
	(0.655)	(0.584)	(0.729)	(0.608)	(0.709)	(0.615)
Pairs	4.621	4.826	4.610	4.792	4.665	4.820
	(0.779)	(0.704)	(0.730)	(0.697)	(0.760)	(0.724)
Qualifying	4.593	4.733	4.534	4.662	4.571	4.678
	(0.738)	(0.683)	(0.748)	(0.725)	(0.765)	(0.728)
Short	4.657	4.965	4.620	4.938	4.659	4.948
	(0.712)	(0.580)	(0.741)	(0.565)	(0.743)	(0.622)
Long	4.963	5.069	4.921	5.047	4.887	5.018
	(0.559)	(0.521)	(0.556)	(0.508)	(0.576)	(0.546)

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

ranked at least once within the top five (or ten) in the past four years in World Figure Skating Championships and other things are equal, her score is higher by 0.35 (or 1.04). The gains come a little bit more from technical scores, which again shows that judges do not differentiate performances in artistic scores. Table 2 also shows that there exists nationalistic bias in figure skating judging. Note that the skater- and judge-country fixed effects are included to allow for the possibility that more judging slots are allocated to countries with better skaters. Even after controlling for the country-specific effects, I find that judges favor compatriot skaters by about 0.18. It is comparable to the estimate of Zitzewitz (2006), 0.17, and slightly larger than 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 free program are higher than those in the short or preliminary round. Scores also get higher as the competition progresses (as starting order increases). This last finding reflects the fact that skaters are seeded in the free program. The starting order is randomly assigned to each skater except in the case of long programs (or free skating programs), where the order is determined by rankings from previous programs. Table 2 also divides total scores into technical and artistic points. One notable thing is that the nationalistic bias is more attributable to the bias in

TABLE 2: Determination of Scores in Level[1]

	Total	Technical	Artistic
Top Five	0.3500	0.1918	0.1582
	(0.0348)	(0.0197)	(0.0166)
Top Ten	1.0387	0.5273	0.5114
	(0.0312)	(0.0176)	(0.0150)
Compatriot Judge	0.1795	0.0762	0.1033
	(0.0502)	(0.0283)	(0.0237)
Female Judge	0.1077	0.0540	0.0536
	(0.0291)	(0.0164)	(0.0138)
Ladies	−0.3654	−0.2056	−0.1599
	(0.0377)	(0.0213)	(0.0178)
Pairs	−0.5312	−0.2756	−0.2556
	(0.0329)	(0.0185)	(0.0156)
Short Program	0.0112	−0.0930	0.1042
	(0.0290)	(0.0165)	(0.0137)
Free Program	0.3168	0.1590	0.1577
	(0.0282)	(0.0155)	(0.0135)
Starting Order	0.0257	0.0139	0.0118
	(0.0017)	(0.0010)	(0.0008)
Constant	8.0258	3.9355	4.0903
	(0.0862)	(0.0475)	(0.0560)
Skater Country Fixed Effect	YES	YES	YES
Judge Country Fixed Effect	YES	YES	YES
$R^2 =$	0.6484	0.6130	0.6558

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

Artistic Score, which makes sense since there should be more room for maneuver. Lastly, compared to scores in the preliminary round, technical scores in the short program are lower while artistic scores are higher. This may be caused by different scoring guidelines for different programs. For example, the short program includes some required elements on which observations are more accurate and point deductions might be made more easily.

4. PREVIOUS OUTLYING SCORES

In this section I examine whether judges submit scores in a strategic fashion, particularly whether they avoid outlying scores when they have already submitted outlying scores for previous skaters. The rules I discussed before show that they should avoid extraordinary deviation from the other judges' scores and, in particular, repeated deviations. The basic specification is therefore dynamic:

$$D_{ij,p} = \alpha D_{ij,p-1} + \beta \overline{D}_{ij}^{p-2} + \gamma Q_{ij,p} + \delta p + \omega_{i,p} + \lambda_{ij} + \nu_{ij,p}, \quad (1)$$

where $D_{ij,p}$ represents judge j's squared deviation for a skater whose starting order is p in year/event/program i. That is, $D_{ij,p} = (S_{ij,p} - \bar{S}_{i,p})^2$ where S denotes score and $\bar{S}_{I,p}$ is the average score of judges other than j.[3] For simplicity I consider only an average cumulative effect of deviations up to the $(p-2)$-th skater by including $\overline{D}_{ij}^{p-2} = \frac{1}{p-2}(\Sigma_{k=1}^{p-2} D_{ij,k})$. I distinguish the immediately preceding deviation from the other previous deviations since the rules discussed before imply that judges should attempt to avoid repeated deviations harder.[4] The coefficients, α and β, are of our major interests. I expect that their signs will be negative in presence of outlier aversion.

Some control variables are included. $Q_{ij,p}$ is a vector of the skater's quality and a constant term. It includes dummy variables that are equal to 1 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 0 otherwise. Starting order, p, is included in case there exists any related systematic effect. $\omega_{i,p}$ represents the performance-specific fixed effect, which is controlled for since there might be some performances that are by nature noisier than others and all judges deviate from each other.

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

Lastly, λ_{ij} represents an unobserved judge-specific skater-invariant effect that allows for heterogeneity across judges. There are two interpretations for this term.[5] First, the effect may represent an individual judge's risk aversion that affects his or her aversion to outlying scores. Judges might as well be heterogeneous in their career concerns. For example, those judges who like to pursue their career as judges are more likely to be conservative and would be more averse to outlying scores. And some judges are more vigorous in their opinions and do not care about deviations.

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, in principle, absolute values of scores do not matter much.[6] 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 scores will be systematically higher or lower and, therefore, his or her 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. I must allow for the individual-specific intercepts of deviations in order to test for outlier aversion.

Unless the distribution of λ_{ij} is degenerate, the lagged dependent variable in equation (1) is necessarily endogenous. Fortunately it is possible to estimate consistently in two steps: (i) eliminate λ_{ij} by the first-differencing transformation and (ii) use the values of the dependent variable lagged by two skaters or more as instrumental variables. This is the Arellano-Bond GMM estimator or the GMM-DIF estimator (Arellano & Bond, 1991). Specifically, I will estimate the following first-differenced equation:

$$\Delta D_{ij,p} = \alpha \Delta D_{ij,p-1} + \beta \Delta \overline{D}_{ij}^{p-2} + \gamma \Delta D_{ij,p} + \delta + \Delta \omega_{i,p} + \Delta \nu_{ij,p}, \tag{2}$$

where Δ represents the first-differencing transformation. I assume that $Q_{j,p}$ is strictly exogenous and $\overline{D}_j^{p-2}$ is predetermined.[7] The key identifying assumption is that the lagged level $D_{ij,p-k}$ will be uncorrelated with $\Delta \nu_{ij,p}$ for $k \geq 2$.

Before progressing further, one might suspect that the squared deviation, although convenient for analysis, is really not what judges are concerned about. Fortunately, it is possible to conduct a direct test of whether $D_{ij,p}$ is meaningful for the judge-assessment system and judges' career concerns. I estimate a simple Probit model that regresses $R_{j,2002}$ on $D_{ij,p,2001}$, where $R_{j,2002}$ is a dummy variable that is 1 if the judge j is reselected for the 2002 championships conditional on the fact that the judge is selected in 2001, and 0 otherwise. Judges' nationality is controlled for to take into account that each national federation has its own unique procedure of recommending judges to the ISU. After controlling for the judge country-specific effect, in Table 3, I find that an increase in the average degree of squared deviation per performance (about 0.13) significantly reduces the probability of reselection by about 0.9 to 2 percent. 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 about 18 to 40 percent. It is obvious that volunteer judges are honored to be selected for major international competitions, like the World Figure Skating Championships. The squared deviation should be one of the important statistics in the judge-assessment system that judges are concerned about.[8]

Table 4 presents our main results. Note that the performance fixed effect is included in addition to the judge-specific fixed effect to control for the possibility that some performances are noisier than others and, for those performances, all judges deviate from each other. And, for comparison, I juxtapose results from the ordinary least squares (OLS), within-group (WG), and Arellano-Bond dynamic panel data (GMM) estimations.[9] Even though the model can be consistently estimated only by GMM, the comparison with these potentially inconsistent

TABLE 3: Judge Selection: Prohibit Model Estimation[1]
(The dependent variable is whether judges were reselected in 2002)

	(1)	(2)	(3)	(4)
$D_{ij,p,2001}$	−0.0713	−0.0800	−0.1579	
	(0.0366)	(0.0365)	(0.0699)	
Positive $D_{ij,p,2001}$				−0.2399
				(0.1155)
Negative $D_{ij,p,2001}$				−0.0935
				(0.0841)
Female Judge		−0.0746	−0.3757	−0.3794
		(0.0183)	(0.0488)	(0.0492)
Short Program		−0.0534	0.1178	0.1185
		(0.0212)	(0.0369)	(0.0369)
Free Program		−0.0243	0.0128	0.0109
		(0.0229)	(0.0409)	(0.0408)
Judge-Country Fixed Effect	NO	NO	YES	YES
Pseudo R^2 =	0.0011	0.0077	0.1117	0.1126

[1]Marginal effects are calculated at sample means. Robust standard errors are displayed in parentheses.

estimates is 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, & Windmeijer, 2000).[10] 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 & Trognon, 1997).

Table 4 shows 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. Also the bias in the WG estimates is small relative to that of the OLS estimates. This finding also accords with 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 $D_{ij,p-1} - \frac{1}{P}(D_{ij,1} + \cdots + D_{ij,p} + \cdots + D_{ij,P})$, and the error term becomes $\upsilon_{ij,p} - \frac{1}{P}(\upsilon_{ij,1} + \cdots + \upsilon_{ij,p-1} + \cdots + \upsilon_{ij,P})$. These two are obviously correlated, above all because $D_{j,p-1}$ and $\frac{1}{P}\upsilon_{ij,p-1}$ are correlated, $\frac{1}{P}D_{ij,p}$ and $V_{j,p}$ 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.

TABLE 4: Previous Deviations: Dynamic Panel Data Estimation[1]

	OLS			WG			GMM		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$D_{ij,p-1}$	0.0367	0.0367	0.0366	−0.0882	−0.0882	−0.0884	−0.376	−0.0376	−0.0373
	(0.0206)	(0.0206)	(0.0206)	(0.0158)	(0.0158)	(0.0158)	(0.0184)	(0.0184)	(0.0185)
$\bar{D}_{ij}^{p-2}$	0.2332	0.2332		−0.7440	−0.7440		−0.1910	−0.1910	
	(0.0496)	(0.0496)		(0.0645)	(0.0645)		(0.0857)	(0.0857)	
Positive $\bar{D}_{ij}^{p-2}$			0.2544			−0.7163			−0.1658
			(0.0691)			(0.0914)			(0.1459)
Negative $\bar{D}_{ij}^{p-2}$			0.2238			−0.7602			−0.2011
			(0.0524)			(0.0747)			(0.0984)
Top Five		0.3781	0.3778		0.0296	0.0284		−0.0348	0.0453
		(0.2721)	(0.2727)		(0.1070)	(0.1070)		(0.0962)	(0.2018)
Top Ten		−0.3481	−0.3478		−0.0569	−0.0565		−0.1262	−0.1663
		(0.2451)	(0.2457)		(0.1023)	(0.1023)		(0.0970)	(0.1231)
Starting Order		0.0083	0.0083		0.0063	0.0063			
		(0.0099)	(0.0099)		(0.0032)	(0.0032)			
Constant	0.0343	0.1157	0.1158	0.2467	0.1347	0.1407	0.3168	−0.0001	−0.0024
	(0.0501)	(0.0546)	(0.0546)	(0.0188)	(0.0736)	(0.0683)	(0.0282)	(0.0030)	(0.0063)
Number of Observations	4,782	4,782	4,728	4,782	4,782	4,782	4,540	4,540	4,540
Number of Judges	242	242	242	242	242	242	235	235	235
Performance Fixed Effect	YES	YES	YES	YES	YES	YES	YES	YES	YES
Judge Fixed Effect	NO	NO	NO	YES	YES	YES	YES	YES	YES
AR(1) Test							−36.91	−36.91	−36.82
AR(2) Test							0.65	0.65	0.64

[1]Robust standard errors are displayed in parentheses. The estimates in columns 7–9 are generated by Arellano-Bond one-step difference GMM. AR(1) and AR(2) tests are the Arellano-Bond tests of first order and second order autocorrelation in the error term.

Finally, the assumption of no serial correlation of $V_{ij,p}$ cannot be rejected. The last two rows in the tables present the Arellano-Bond test statistics for autocorrelation. 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, and the AR(2) structure is rejected across the board. Also the validity of the instruments is supported by the Sargan test of over-identifying restrictions.

The GMM estimates imply that the deviation of a judge's score for the previous skater significantly decreases the deviation for the current skater. The existence of outlier aversion makes judgments biased and, in particular, is harmful to those skaters who did an unusually good job. The order of appearance in a competition matters (Ginsburgh & Van Ours, 2003). For example, suppose that the judge's score is deviated from the average of the other judges' by the extent of 0.45.[11] The estimates here imply that the deviation pressures judges to be biased by about 0.09 of a point $(=\sqrt{\alpha \times 0.45^2} = \sqrt{0.038 \times 0.45^2})$ toward the average for the current player. I also estimate each program separately. The results are consistent. Similarly, the degree of outlier aversion 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.

I also estimate the effects of positive and negative deviations separately in column 9. I expect that judges should be more averse to positive extreme scores since a positive bias is considered favoritism, which is a more sensitive issue in this sport than inaccurate scoring. However, I find that judges are equally responsive to positive and negative deviations. The effects are not statistically different.

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 (the unit of score). 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.

The marginal effect of the average squared deviation up to the $(p-2)$-th previous performance is larger in size than that of the one-time deviation for the immediately preceding performance. When skaters' quality is controlled, β (-0.1910 in column 8) is five times as large as α (-0.0376) in absolute terms. This seems reasonable, because β picks up a kind of cumulative effect of α.[12] For example, the magnitude of β 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 about 0.2 $(=\sqrt{0.1910 \times 0.45^2})$.

Other results are also consistent with prior expectation. Judges are more in agreement for top skaters, although the effects are marginally significant. 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 scoring top-ranked skaters.[13] Table 5 shows the results when we include only the programs with randomly selected orders. The results are very similar as before. The effects of previous deviations are even stronger.

5. INTERIM JUDGING SYSTEM

In this section, I exploit a natural experiment of the judging system reform. In 2002 the ISU adopted a new system, the Interim Judging System, in which judges' names are concealed on the scoreboard from outside observers, including judges themselves. In addition, the scores of only a subset of the judges, chosen at random, are used to arrive at the final ranking. The new system was first implemented in 2003.

The change in the judging system provides another opportunity to test the existence of outlier aversion, since one might expect that judges would be less pressured to agree under the new system. The ISU itself states "anonymity reduces

TABLE 5: Events with Randomly Selected Orders[1]

	(1)	(2)	(3)
$D_{ij,p-1}$	−0.0504	−0.0504	−0.0501
	(0.0217)	(0.0217)	(0.0217)
$\bar{D}_{ij}^{p-2}$	−0.2829	−0.2829	
	(0.1035)	(0.1035)	
Positive $\bar{D}_{ij}^{p-2}$			−0.2582
			(0.1762)
Negative $\bar{D}_{ij}^{p-2}$			−0.2918
			(0.1170)
Top Five		−0.0276	−0.2012
		(0.1223)	(0.1188)
Top Ten		−0.2622	0.1666
		(0.1161)	(0.1748)
Constant	−0.0124	−0.0001	−0.0090
	(0.0059)	(0.0030)	(0.0068)
Performance Fixed Effect	YES	YES	YES
Judge Fixed Effect	YES	YES	YES
AR(1) Test	−31.41	−31.41	−31.33
AR(2) Test	0.25	0.25	0.24

[1]Robust standard errors are displayed in parentheses. The estimates are generated by Arellano-Bond one-step difference GMM. AR(1) and AR(2) tests are the Arellano-Bond tests of first order and second order autocorrelation in the error term.

the risk of judges coming under external pressure" (International Skating Union, 2002). The "external pressure" referred to by the ISU is meant to be nationalistic favoritism. Besides that, anonymity will relieve 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. Thus it is reasonable to suppose that the introduction of anonymity, though it cannot remove completely, significantly weakens judges' incentives for outlier aversion.

Table 6 shows the simple mean comparisons of deviations before and after the system changed. For robustness, I use three measures of score dispersion for a skater p:

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

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

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

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.

As shown in Table 6, 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.

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. To meet this objection, I 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

TABLE 6: Dispersion: Before-and-After Comparison[1]

	Technical			Artistic		
Men	Before	After	Δ	Before	After	Δ
Standard Deviation	0.1589	0.1825	0.0236	0.1629	0.1698	0.0068
	[0.0657]	[0.0753]	(0.0087)	[0.0663]	[0.0698]	(0.0084)
Absolute Deviation	0.1793	0.2056	0.0263	0.1836	0.1913	0.0077
	[0.0765]	[0.0889]	(0.0102)	[0.0793]	[0.0799]	(0.0100)
Range	0.4651	0.5989	0.1339	0.4672	0.5553	0.0881
	[0.2007]	[0.2674]	(0.0284)	[0.2015]	[0.2308]	(0.0267)
Ladies						
Standard Deviation	0.1766	0.2069	0.0303	0.1785	0.1986	0.0201
	[0.0860]	[0.0723]	(0.0102)	[0.0781]	[0.0806]	(0.0099)
Absolute Deviation	0.1995	0.2338	0.0342	0.2022	0.2238	0.0216
	[0.1010]	[0.0841]	(0.0120)	[0.0925]	[0.0938]	(0.0116)
Range	0.5192	0.6740	0.1548	0.5192	0.6479	0.1287
	[0.2644]	[0.2556]	(0.0327)	[0.2318]	[0.2660]	(0.0304)
Pairs						
Standard Deviation	0.1748	0.1770	0.0021	0.1638	0.1837	0.0199
	[0.0948]	[0.0818]	(0.0097)	[0.0951]	[0.0923]	(0.0101)
Absolute Deviation	0.1969	0.1999	0.0030	0.1835	0.2051	0.0216
	[0.1124]	[0.0961]	(0.0115)	[0.1122]	[0.1063]	[0.0118]
Range	0.5095	0.5805	0.0710	0.4721	0.6008	0.1287
	[0.2771]	[0.2672]	(0.0292)	[0.2817]	[0.3050]	(0.0307)

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

$$\chi_p = b + \beta_1 B_p + \beta_2 A_p + \beta_3 B_p A_p + Q_{p\gamma} + X_p \delta + u_p, \quad (3)$$

where ξ_p is one of the three dispersion measures.[14] A_p is the indicator of anonymity (one for the new system and zero for the old system). Since the Interim Judging System was first implemented in 2003, this variable is a yearly dummy variable with 1 for 2003 and 0 for 2001 and 2002. X_p is a vector of indicators for events and programs. It also includes a yearly dummy for 2002, which is included for generality, although there is no systematic change between 2001 and 2002. Furthermore, I include the skater-country fixed effect to control for any country-specific unobserved heterogeneity. Note that the judge-country fixed effect cannot be included in the judge panel-level analysis. Q_p is a vector of measures of skaters' quality.

Table 7 shows that scores are more dispersed after the introduction of anonymity and random selection, even after controlling for the nationalistic bias. The standard deviation significantly increases by about 0.015, the absolute deviation increases by about 0.014, and the range increases by about 0.096. The magnitude ranges from 14 to 37 percent of one standard deviation of each measure.[15]

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

6. DISCUSSION

The purpose of this paper is to illustrate that subjective evaluators are sensitive to the incentive structure they work within and that their reports may depend upon how they are monitored and assessed. I focus on outlier aversion 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. Judges tend to rank skaters in accordance with pre-performance information, putting less weight on the actual competition.

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

TABLE 7: Dispersion: Regression Analysis[1]

			Range = Max − Min		
	Standard Deviation	Absolute Deviation		Max	Min
Year 2002	0.0001	−0.0012	−0.0041	−0.0254	−0.0213
	(0.0054)	(0.0064)	(0.0162)	(0.0272)	(0.0340)
Year 2003 (Anonymity)	0.0150	0.0140	0.0963	−0.0405	−0.1368
	(0.0053)	(0.0063)	(0.0170)	(0.0296)	(0.0356)
Compatriot	0.0012	0.0010	−0.0131	0.0187	0.0318
	(0.0062)	(0.0073)	(0.0180)	(0.0309)	(0.0384)
Year 2002 × Compatriot	−0.0008	−0.0006	0.0117	0.0021	−0.0096
	(0.0084)	(0.0099)	(0.0251)	(0.0433)	(0.0547)
Year 2003 (Anonymity) × Compatriot	0.0077	0.0107	0.0524	0.1066	0.0543
	(0.0082)	(0.0097)	(0.0260)	(0.0442)	(0.0538)
Top Five	−0.0299	−0.0356	−0.1027	0.1741	0.2767
	(0.0051)	(0.0060)	(0.0163)	(0.0246)	(0.0335)
Top Ten	−0.0345	−0.0406	−0.1096	0.4743	0.5839
	(0.0047)	(0.0055)	(0.0148)	(0.0235)	(0.0304)
Ladies	0.0205	0.0244	0.0660	−0.1175	−0.1835
	(0.0043)	(0.0050)	(0.0136)	(0.0259)	(0.0307)
Pairs	0.0125	0.0144	0.0363	−0.2582	−0.2945
	(0.0043)	(0.0051)	(0.0132)	(0.0224)	(0.0275)
Qualify	0.0445	0.0545	0.0955	−0.0960	−0.1915
	(0.0039)	(0.0046)	(0.0123)	(0.0210)	(0.0260)
Short	0.0237	0.0279	0.0732	−0.0622	−0.1354
	(0.0035)	(0.0042)	(0.0118)	(0.0208)	(0.0254)
Artistic	−0.0035	−0.0045	−0.0153	0.1734	0.1887
	(0.0032)	(0.0038)	(0.0101)	(0.0172)	(0.0212)
Constant	0.1686	0.1922	0.5275	4.5255	3.9980
	(0.0073)	(0.0086)	(0.0236)	(0.0449)	(0.0527)
Skater Country Fixed Effect	YES	YES	YES	YES	YES
R^2 =	0.2700	0.2724	0.2757	0.6159	0.6168

[1]N = 2,020. Robust standard errors are displayed in parentheses.

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 ways to aggregate different preferences and minimize idiosyncratic errors in subjective evaluation. My findings are consistent with recent applied psychology and management literature regarding performance appraisals that finds raters pursue their own objectives and that it is important to understand rater motivation (Murphy, Cleveland, Skattebo, & Kinney, 2004; Murphy & De Schon, 2000).

The results also imply that agreement among evaluators is not always desirable. Firms often use subjective evaluation in group decision-making processes. Unfortunately, as the findings have demonstrated, these processes are subject to outlier aversion 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. AQ4

NOTES

1. This is conceptually related to conformity (Bernheim, 1994) and "Groupthink" (Janis, 1983). Janis documents historical moments such as the Cuban missile crisis and Korean War where conforming to group norms within the president's inner circle led to disastrous consequences. Similarly there have been many studies showing that macroeconomic forecasters and equity analysts strategically herd. See, for example, Lamont (2002) and Welch (2000). AQ2

2. Recently there have been a few studies that use sports data to test theories regarding subjectivity of performance evaluation. For example, refer to Garicano, Palcios, and Prendergast (2005), Campbell and Galbraith (1996), and Zitzewitz (2006).

3. The convexity of the measure of deviation seems to be reasonable since judges should avoid extreme scores. The choice of the measure will be also empirically rationalized later. See Table 3 and related discussion.

4. The lag structure will be tested later.

5. I assume that a judge may have different judge-specific effects in different programs within

an event. 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.

6. One might expect that what judges really care about should be their ranking, not scores. However, the ISU explicitly mentions that they investigate the marks awarded (absolute scores).

7. The lagged values of $\bar{D}_j^{p-2}$ can be used as additional instruments. However it is not feasible in this paper because of multicollinearity with the performance fixed effect.

8. It is interesting to see that female judges are less likely to be reappointed. And it is also found that the probability of reappointment depends more on positive deviations.

9. The WG estimator is the OLS estimator with the judge fixed effect. The fixed effect is removed by within-group transformation.

10. The WG estimator eliminates λ_p by transforming the original observations in to deviations from individual means. However this transformation induces a nonnegligible correlation between the transformed lagged dependent variable and the transformed error term.

11. It is the average of the distance between median and extrema.

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

13. I reestimated the GMM model on technical and artistic scores, separately. I found that artistic scores are indeed more responsive to previous deviations. Again, it implies that judges can manipulate artistic score more easily than technical score.

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

15. 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.

REFERENCES

Arellano, Manuel, & Bond, Stephen (1991). Some tests of specification for panel data: Monte Carlo Evidence and an Application to Employment Equation. *Review of Economic Studies*, *58*, 277–297.

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

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

Blundell, Richard, Bond, Stephen, & Windmeijer, Frank (2000). Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimators. In Badi Baltagi (Ed.), *Advances in Econometrics* (Vol. 15). Amsterdam: JAI Elsevier Science, 2000.

Bond, Stephen (2002). Dynamic panel data models: A guide to micro data methods and practice. *Portugese Economic Journal*, *1*, 141–162.

Borman, Walter C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology*, *67*, 3–9.

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

Garicano, Luis, Palacios, Ignacio, & Prendergast, Canice (2005, May). Favoritism under Social Pressure. *Review of Economics and Statistics*, *87*(2), 208–216.

Ginsburgh, Victor A., & van Ours, Jan C. (2003, March). Expert opinion and compensation: Evidence from a musical competition. *American Economic Review*, *93*(1), 289–296.

Ilgen, Daniel R., Barnes-Farrell, Janet L., & McKellin, David B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes*, *54*, 321–368.

International Skating Union (2002, December 27). Year-end update: Figure skating judging for the 2002/3 season and beyond. Media Release.

International Skating Union (1999). Sanctions and Penalties. *ISU Communication*, *1025*.

Janis, Irving L. (1983). *Groupthink: Psychological studies of policy decisions and fiascoes* (2nd ed.). Boston: Houghton Miffin Company.

Lamont, Owen A. (2002, July). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior and Organization*, *48*(3), 265–280.

Milkovich, George T., & Wigdor, Alexandra K. (Eds.). (1991). *Pay for performance: Evaluating performance appraisal and merit pay.* Washington DC: National Academy Press.

Murphy, Kevin R., & Cleveland, Jeanette N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspective.* Sage Publications. AQ5

Murphy, Kevin R., & Cleveland, Jeanette N. (1991). *Performance appraisal: An organizational perspective.* MA: Allyn and Bacon, 1991. AQ6

Murphy, Kevin R., & De Schon, Richard (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.

Murphy, Kevin R., Cleveland, Jeanette N, Skattebo, Amie L., & Kinney, Ted B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology*, *89*(1), 158–164.

Nickell, Stephen J. (1981). Biases in dynamic models with fixed effects. *Econometrica*, *49*(6), 1417–1426.

Prendergast, Canice (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*(1), 7–63.

Prendergast, Canice (1993, September). A theory of "yes men." *American Economic Review*, *83*(4), 757–770.

Saal, Frank, Downey, Ronald, & Lahey, Mary Anne (1980, September). Rating the ratings: Assessing the quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Saxonhouse, Gary R. (1976, March). Estimated parameters as dependent variables. *American Economic Review*, *66*(1), 178–183.

Sevestre, Patrick, & Trognon, Alain (1996). Dynamic linear models. In *The econometrics of panel data.* Boston and London: Kluwer Academic.

Topel, Robert, & Prendergast, Canice (1996, October). Favoritism in Organizations. *Journal of Political Economy*, *104*(5), 958–978.

Viswesvaran, Chockalingam, Ones, Deniz S., & Schmidt, Frank L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*(5), 557–574.

Weekley, Jeff A., & Gier, Joseph A. (1989). Ceilings in the reliability and validity of performance ratings: The case of expert raters. *Acedemy of Management Journal*, *32*(1), 213–222.

Welch, Ivo (2000). Herding among security analysts. *Journal of Financial Economics*, *58*, 369–396.

Yamaguchi, Kristi, Ness, Christy, & Meacham, Jody (1997). *Figure skating for dummies.* Foster City, CA: IDG Books.

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

AUTHOR QUERIES

RE "Outlier Averision,"

AQ1: Please provide author name and affiliation.

AQ2: The Cuban missile crisis is widely considered an example of group reasoning at its best and as having averted disaster. Recommend using another example.

AQ3: Recommend changing 'ladies' to 'women' (to avoid gender bias) unless this is the term used by the skating authorities.

AQ4: Please include a brief author bio.

AQ5: Please provide the city of publication.

AQ6: Please provide the city of publication.