NBER WORKING PAPER SERIES

DOES TRANSPARENCY REDUCE FAVORITISM AND CORRUPTION? EVIDENCE FROM THE REFORM OF FIGURE SKATING JUDGING

Eric Zitzewitz

Working Paper 17732
http://www.nber.org/papers/w17732

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2012

The author would like to thank Jay Emerson, Ray Fisman, Leo Kahane, two anonymous referees, and anonymous former figure skating officials for helpful comments. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Does Transparency Reduce Favoritism and Corruption? Evidence from the Reform of Figure Skating Judging

Eric Zitzewitz

NBER Working Paper No. 17732
January 2012
JEL No. D7,D8

ABSTRACT

Transparency is usually thought to reduce favoritism and corruption by facilitating monitoring by outsiders, but there is concern it can have the perverse effect of facilitating collusion by insiders. In response to vote trading scandals in the 1998 and 2002 Olympics, the International Skating Union (ISU) introduced a number of changes to its judging system, including obscuring which judge issued which mark. The stated intent was to disrupt collusion by groups of judges, but this change also frustrates most attempts by outsiders to monitor judge behavior. I find that the "compatriot-judge effect", which aggregates favoritism (nationalistic bias from own-country judges) and corruption (vote trading), actually increased slightly after the reforms.

Eric Zitzewitz
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
eric.zitzewitz@dartmouth.edu

Does Transparency Reduce Favoritism and Corruption? Evidence from the Reform of Figure Skating Judging

In most principal-agent relationships, transparency is thought to improve social welfare by facilitating monitoring, reducing moral hazard problems. In some cases though, transparency may have perverse effects that make it less socially desirable.

For example, transparency may facilitate collusion by making cheating on a collusive agreement easier to detect. A lengthy literature in industrial organization discusses the role transparency can play in softening competition (e.g., Stigler, 1964; Green and Porter, 1984). Industries commonly mentioned as exemplifying this effect include airlines, internet retailing, retail gasoline, and financial market making. In agency settings, transparency can facilitate monitoring by third parties as well as by agents' principals. For example, it was argued that requiring mutual funds to disclose their votes on corporate proxies would lead to more pro-management voting, since disclosure would allow management to punish fund sponsors who voted against them, such as by withholding 401k or underwriting business (Davis and Kim, 2007). Voice votes in legislatures are often thought to facilitate influence by special interests by hampering constituents' monitoring (e.g., Arnold, 1993), but they are also defended as allowing legislators to act in the public interest and evade monitoring by interest groups. The same tradeoffs arise with hiring and promotion decisions, which are typically made by secret ballot in academia, but not in for-profit firms. Several academic departments I am familiar with have debated whether open voting or secret ballots are optimal. Some argue that open voting would restrain personal and field-related biases, but others argue that it would perversely encourage untenured faculty to mimic the biases of their senior colleagues (as in Prendergast, 1993).

In short, it is difficult to determine on purely theoretical grounds whether transparency will generate better outcomes. This motivates turning to empirical analyses of settings where one can test how the advantages and downsides of transparency net.

This paper examines one such setting. While the general trend in society is arguably toward greater transparency, this paper examines a relatively unique policy change that significantly *reduced* transparency -- with the stated goal reducing favoritism and corruption. Following vote trading scandals in the 1998 and 2002 Olympics, the International Skating Union (ISU) introduced a number of reforms to its judging and scoring system. The most controversial was no longer reporting which judge gave which score. The ISU's rationale was that anonymity would reduce outside pressures on judges, such as those experienced by the French judge in the 2002 Olympics Pairs competition, who was reportedly pressured by some combination of her national federation and the Russian mafia to vote for a Russian pair in exchange for a Russian vote for a French couple in Ice Dancing.[2] Opponents of anonymity argued that it would also reduce the potential for outside monitoring. The United States Figure Skating association opposed the change on these grounds, and, while it has adopted most elements of the new ISU system, it continues to disclose which judge gave which scores at U.S. national competitions.[3]

Anonymity frustrates the most straightforward approach to studying judging biases in figure skating, which is to compare different judges' scoring of the same performance (Seltzer and Glass, 1991; Campbell and Galbraith, 1996; Sala, Scott, and Spriggs, 2007). But unlike many low-transparency settings, such as in the mutual fund, voice voting, and secret ballot examples discussed above, in this case, anonymity does not completely frustrate the measure of bias.

As I found in Zitzewitz (2006), having a compatriot on the judging panel yields both a higher score from the compatriot judge and higher scores from other judges (relative to the scores they give the same competitors when they do not have a compatriot judge). The former difference can be attributed to nationalistic bias, while latter difference may reflect vote trading. When there is vote

[2] For allegations of influence from the French federation, see e.g. Garrahan, Matthew, "Olympic Committee Awards a Second Gold Medal," *Financial Times*, 2/15/2002. For allegations of Russian mafia involvement see, e.g., Appleson, Gail, "Man Arrested on Charges of Fixing Olympic Event," *Reuters News*, 7/31/2002.

[3] Russia has more recently also expressed opposition to judge anonymity. See, for example, *REGNUM News Agency*, "Russia to Suggest Reform of Figure Skating Judging," 6/21/2006.

trading, the within-performance comparison approach taken in past work actually significantly understates the total advantage a skater gains from a compatriot judge, as it nets out the effects of vote trading on other-country judges' scores, when these effects should be added instead.

The size of the total effect of a having compatriot judge (including both nationalistic bias and vote trading) is what is most relevant for the fairness of the competition. Fortunately, judge anonymity does not prevent me from estimating this combined effect -- it merely prevents me from decomposing it into nationalism and vote trading. When I compare the total compatriot-judge effect before and after judges' scores were anonymized immediately after the 2002 scandal, I find it increased by about 20 percent (although this increase was not statistically significant). About a year after the scandal, the ISU introduced a new, more complicated, scoring system. The new system significantly reduced the role of judges' subjective scores and also, by virtue of its complexity, made the role of individual judges less salient, which can arguably be considered a further reduction in transparency. Despite the fact that the lower weight given subjective scores should have decreased incentives for bias, the compatriot-judge bias again increased slightly. Taken together, the results suggest little evidence that reducing transparency achieved its goal of reducing favoritism and corruption. If anything, the judging reforms were follows by modest increases in bias.

The remainder of the paper is organized as follows. The next section provides background on the pre-and post-2002 judging systems and discusses related academic work. The third section describes the pre and post-2002 judging data. The fourth section presents the methodology for estimating compatriot-judge effects and presents the main results. A fifth section presents additional evidence aimed at understanding how the nationalistic bias and vote trading components have changed post-reform. A sixth section discusses possible alternative explanations for the results. A conclusion follows.

II. Background

Sports are often a useful setting in which to examine phenomena that are of broader significance. Sports provide the repetition and exogenous variation that one can otherwise only achieve in an experiment, but with the advantage of larger sample sizes and participants with strong interest in the outcome. Most related to this paper are studies of bias and collusion in sports. A number of studies have found biases in officiating -- racial biases in basketball (Price and Wolfers, 2010) and baseball (Parsons, et. al., 2011 and Chen, 2009), home-team biases in soccer (Garicano, et. al. 2005; Dohmen, 2005) and basketball (Price, Remer, and Stone, 2012), as well as nationalistic biases in diving (Emerson, Seltzer, and Lin, 2009) and other judged winter sports (Zitzewitz, 2002). Other work has examined collusion among participants or participants and outsiders, such as purposefully losing sumo matches (Duggan and Levitt, 2002), agreeing to draw chess matches to save energy (Moul and Nye, 2009), or colluding with gamblers to avoid covering a point spread (Wolfers, 2006).

Figure skating in particular has also attracted the attention of those interested in how to best aggregate the opinion of potentially biased decision makers. As discussed by Bassett and Persky (1994), Wu and Yang (2004), and Gordon and Truchon (2008), the pre-2002 figure skating judging system was unique among judged sports in using majority rule to determine skaters' placements. Skaters were ranked based on their median rank among the judges: a skater who was ranked first by five of nine judges was the winner, even if the other four judges ranked the skater dead last. Most other judged sports, including figure skating after 2002, aggregate judges' scores using a trimmed mean (i.e. an average with the top and bottom scores excluded), which allows the results to be affected by both the sign and the magnitude of a judge's view on the difference in two competitors' quality. These approaches have different advantages: as Bassett and Persky and Wu and Yang emphasize, the median rank provides strong safeguards against manipulation by a minority; while as Zitzewitz (2006) discusses, this may be at the cost of encouraging manipulation by a majority.

Clearly, these issues also arise in other social choice settings. Part of the motivation for examining them in figure skating is that while biases are often difficult to detect and predict in "real-world" settings, nationalistic bias in figure skating is large and predictable. In Zitzewitz (2006), I find that same-country judges rank competitors 0.45 (within-performance) standard deviations higher than other judges.[4] Such a bias can be detected with statistical confidence in a sample as small as 20-25 performances, which facilitates analysis of how biases vary with scrutiny and incentives in larger samples.

The focus of this paper is on how the post-2002 changes to figure skating judging affected biases in favor of competitors with compatriot judges. Despite the considerable academic interest in figure skating judging, it is to my knowledge the first to examine judging bias in the post-2002 system. I now briefly describe the pre and post-2002 systems. From around 1900 to 2002, judges assigned skaters scores of 0.0 to 6.0 (with a 0.1 minimum increment) for technical merit and artistic impression. Competitors were ranked by each judge based on the total of these scores, and, as discussed above, overall rankings were determined by competitors' median ordinal ranking.[5]

Judging scandals at the 1998 and 2002 Olympics motivated significant changes to this system. In 1998, Ukrainian judge Yuri Balkov was taped by another judge pre-announcing the order in which he planned to rank the contestants.[6] In 2002, a coalition of five judges allegedly pre-agreed to rank a Russian pair first, and they followed through on the agreement despite a general consensus that a Canadian pair had performed better. In her initial comments after the competition (which she later

[4] Campbell and Galbraith (1996) do not report within-performance standard deviations, but they do find similarly sized biases in points for earlier time periods.

[5] In 1998, the system shifted from constructing a single ranking of skaters based on their median ordinal ranking to making pair-wise comparisons for adjacently placed skaters (see Wu and Yang, 2004 and Gordon and Truchon, 2008). In the latter system, a skater could place ahead of another with a better median ranking if she was ranked more highly by a majority of judges. Rules existed to break ties and, in the latter system, resolve non-transitivities.

[6] See, e.g., Crockatt, Joan, "Corruption on Ice: The Sale-Pelletier Scandal is Only the Latest in a Long Line of Bad Judging," *The Gazette (Montreal)*, 2/16/2002, p. B7.

recanted), French judge Marie-Reine Le Gougne indicated that she had been pressured by her national federation to vote for the Russian pair in exchange for a vote for a French couple in Ice Dancing.

These scandals motivated changes aimed at reducing judges' discretion by making scoring more objective and reducing the potential for coalitions to pre-determine results. A more complicated scoring system, called the "code of points", was introduced. The new system is most analogous to that used in diving, in that it combines a pre-determined "base value" based on the difficulty of the program with subjective scores for the execution quality of each technical element and for up to five aspects of the program's overall quality.[7] Each competitor provides a list of the technical elements (e.g., jumps, spins) that a program will include, and these elements have predetermined point values. A technical panel, usually composed of a technical controller, a technical specialist, and an assistant technical specialist, reviews the program, determines whether all elements were executed as planned, revises the base value if needed (e.g., if a double jump was substituted for a triple), and identifies both obvious issues for which deductions must be assessed (e.g., falls) and non-obvious issues for the judges' consideration (e.g., a takeoff that may have been from the incorrect skate edge). The judges then assign each element a grade of execution (GOE) ranging from +3 to -3 and give scores for up to five components of a program's overall quality. For almost all events, the five components of overall quality are skating skills, transitions/linking footwork, performance/execution, choreography/composition, and interpretation/timing.[8] Judges give scores between 0.00 and 10.00 (minimum increment 0.25) for each

[7] The ISU's description of the new system is available at: http://www.isu.org/vsite/vcontent/page/custom/0,8510,4844-152094-169310-31825-132302-custom-item,00.html and the U.S. Figure Skating Association's description is available at http://www.usfigureskating.org/New_Judging.asp?id=289 (all websites last accessed February 1, 2010).

[8] For ice dancing compulsories, there are four components (skating skills, timing, performance/execution, and interpretation).

of these components. The execution grades are converted into points using a "scale of values" table; and the program components are multiplied by factors that differ for different disciplines and rounds.[9]

Unlike in the "6.0" system, in the code of points system, scores are first aggregated across judges for each element and component and then the scores for the elements and components are added to determine overall placement. The aggregation method was also changed from median ranking to a trimmed mean. Controversially, before the trimmed mean is calculated, a predetermined number of scores are dropped at random (e.g., when there are 12 judges, 3 are dropped at random and the trimmed mean of the remaining 9 is used). As Emerson (2007) discusses, this randomness adds noise to results; he cites the example of the bronze and silver medalists in the 2006 World Championship Pairs competition, whose positions would have been reversed had scores been calculated without randomly dropping judges.[10]

In addition to these changes to the scoring system and aggregation, the ISU also stopped disclosing which judges gave which scores. In addition, at many events, judges and the technical committee are no longer identified as representing a country, but are instead identified only as representing the ISU. These changes seemed designed to downplay the role of individual judges, and especially their nationality, in determining the results.

The most controversial of these changes has been the anonymization of the judges and the random selection of which marks to count. The ISU defends both changes as helping to deter collusion by a coalition of judges. It argues that anonymization allows judges to secretly defect against a collusive agreement, while randomization increases the size of a coalition needed to affect results with certainty. Opponents of anonymization (e.g., skatefair.org) have argued that it reduces public scrutiny of judges'

---

[9] See http://www.usfigureskating.org/content/ISUCommunication1400.pdf for an examples of a scale of values table, and http://www.usfigureskating.org/New_Judging.asp?id=289 for the factors applied to component scores.

[10] As Emerson and Arnold (2011) discuss, in 2010 the ISU began frustrating analyses of the impact of randomization by shuffling the columns in which judges' scores were reported. This change in reporting does not affect my analysis, both because it occurs after the end of my sample and because I use the average pre-randomization score as my outcome of interest.

biases. Critics of randomization have argued that it introduces noise into results (e.g., Emerson, 2007) and that it introduces the potential for manipulation by the designers of the software (e.g., Loosemore, 2002). As mentioned above, the United States Figure Skating Association has declined to adopt anonymization and randomization in its national competitions.

Another change that was discussed, but apparently not implemented, was to have the ISU choose judges rather than the national federations. In Zitzewitz (2006), I found that national federations tended to select judges with the largest past nationalistic biases, while in ski jumping, the central body selected judges and choose those with the smallest past biases. The current system in skating is to have an event's organizer (the ISU for major championships, or the host country for Grand Prix events) select judge countries and for those countries' national federations to select the judges. Rules are in place to limit the number of judging roles given to any judge, and the opportunity to nominate a judge appears to be shared among countries in rough proportion to their representation among competitors.[11]

The changes discussed above were implemented in two phases. Anonymization and randomization were introduced for the 2002-3 season.[12] The code of points was introduced for Grand Prix events only in the 2003-4 season and for all events in the following season. This difference in timing allows one consider the effects of the two changes separately.

III. Data

In this paper, I examine three samples, one prior to the 2002 scandal, one for the events in 2002-4 that retained the 6.0 system but had judge anonymity, and one for events with both the code of points

[11] See, for example, the section "Entry of Judges" on http://www.skatingjapan.jp/InterNational/2008-2009/nhk/index.htm.

[12] In addition, after the 2002 scandal, the masking of judges' countries and of the mapping of judges to scores was retroactively applied to some score sheets (e.g., http://www.icecalc.com/events/owg2002/results/SEG001.HTM). Fortunately, I collected the pre-scandal sample used in this paper before this retroactive masking occurred. In the analysis that follows, I treat pre-scandal events has not having judge anonymity, since the judges were not anonymous at the time of the event.

system and judge anonymity. The pre-scandal sample is the same used in Zitzewitz (2006): it covers 16 events from the 2000-1 and 2001-2 seasons. The "anonymous 6.0" sample includes 23 events from the 2002-3 and 2003-4 seasons. The "code of points" sample includes 107 events from the 2003-4 to 2008-9 seasons. The pre-scandal sample was collected as described in Zitzewitz (2006), while both post-2002 samples were collected from the ISU website.[13]

All three samples include major championships (the Olympics and the European, Four Continents, World and World Junior championships), Grand Prix events, and Junior Grand Prix events. Almost of all these events contain competitions for Men, Ladies, Pairs, and Ice Dancing.[14] Each competition contains multiple rounds; there are always at least two rounds in figure skating (a Short Program and a Free Skate) and almost always at least three rounds in Ice Dancing (a Compulsory Dance, Original Dance, and Free Dance).[15] For a few competitions, such as World Championships and World Junior Championships, there were occasionally additional qualifying or compulsory rounds. A total of 2,976, 3,678, and 15,159 unique performances are available in the pre-scandal, anonymous 6.0, and code of points samples, respectively.

Table 1 provides summary statistics for the three samples. The most notable difference between the samples is the expansion in the size of judging panels immediately after the scandal. Judging panels at major championships were increased from 9 to 14 judges immediately after the scandal, and then reduced to 12 when code of points system was put in place.[16] Due to the increase in the size of judging panels, the share of competitors with a compatriot judge in a given round increased from 53 percent to 69 percent.

---

[13] For the 2002-3 to 2007-8 seasons, these events are listed on http://www.isufs.org/events/. Scores were collected from both the HTML pages and the PDF files available on these sites.

[14] Some junior grand prix events omitted Pairs competitions in the 2006-7, 2007-8, and 2008-9 seasons.

[15] Grand Prix final events have only two rounds in Ice Dancing.

[16] Judging panels were subsequently cut to 9 for major championships in the 2008-9 season as a cost cutting measure (see, Smith, Beverley, "Judging Panels to Shrink at Major Championships, Including Next Olympics; Figure Skating: New Rules Raise Old Concerns About Mark Manipulation," *Globe and Mail*, 10/10/2008, S1).

Table 2 provides summary statistics on scores given to competitors under the 6.0 and code of points systems. The table reports both within-performance standard deviations and within-round (and between-competitor) standard deviations. The latter are about three times as large under the 6.0 system, but only about 1-1.5 times for individual components under the code of points system. This implies that there was more consistency among the judges as to which performances earn the highest marks under the 6.0 system. For scores that are the sum of separately judged components, the standard deviation of the sum is usually only slightly below the sum of the standard deviations of the components, implying that the scores given to a particular performance are very correlated. The exception to this is total element score, which combines an objectively measured base value (i.e., difficulty rating) and subjective grade of execution scores. Base value and grade of execution are essentially uncorrelated across skaters in a given round, implying that competitors who attempt more difficult programs earn similar average execution quality scores.

The summary statistics for the pre-scandal and the anonymous 6.0 sample do not appear very different from one another with the exception of a larger difference between the maximum and median score. This increase could simply be due to the increase in the number of judges, but the increase in the inter-quartile range is not as large, suggesting that it might reflect more extreme judgments from the most favorable judge.[17] We conduct some analysis in Section V to test whether this change is driven by compatriot judges.

[17] While the ratio of the max-median spread and the inter-quartile spread clearly increases to infinity with the number of judges if scores are unbounded, the effects of small changes in the number of judges can actually be non-monotonic. For example, if judges' scores were drawn from identical and independent normal distributions, the max-median to inter-quartile ratio would be 1.30, 1.23, and 1.29 for 9, 12, and 14 judges, respectively. Empirically, this ratio is 1.22 for 9 judge events in the pre-scandal sample, 1.25 for 9 judge events and 1.21 for 14 judge events in the anonymous 6.0 sample, and 1.61 for 9 judge events and 1.48 for 12 judge events in the code of points sample. Increases in the number of judges do not appear to be the source of the higher ratios under the code of points system.

IV. Results

This section analyzes the effect that representation on the judging panel has on a competitor's score. We use the following simple empirical model:

$$s_{crp} = b \cdot j_{cr} + a_c + n_r + e_{crp} \quad (1)$$

where s_{crp} is the score given competitor *c* for performance *p* in round *r*. J_{cr} is an indicator variable equal to one when country *c* has a judge on the panel for round *r*, a_c is a competitor fixed effect that captures the average quality of that competitor's performances, n_r is a round fixed effect that captures the leniency of judging at a particular round, and e_{crp} is a performance specific error term. Note that the round fixed effects subsume any variables that are constant within round, such as the location, date, composition of the judging panel, the discipline (e.g., ice dancing, ladies', pairs, men's), as well as the type of skating done in that round (e.g., compulsories, short program, free skate).

The identifying assumption required to interpret *b* as an estimate of bias in this regression is that the true quality of a given competitor's performances be no higher when a compatriot is on the judging panel. Thus if there are factors that might be correlated with both within-competitor variation in performance quality and panel composition, it is important to include them in the regression. The most obvious such factor is the location of the competition. Competitors might be expected to perform better in their home countries and in locations involving less travel. The hosting federation selects the judge countries, and it might be more likely to include its own judges and to reduce travel costs by selecting judges from nearby countries. I therefore include controls for home-country and home-region events. As discussed below, competitors do receive slightly higher scores when competing in their home countries, but failing to control for this effect does not meaningfully bias the results.

Table 3 presents summary statistics on major countries' share of events hosted, competitor performances, and judging slots, as well as on how judging slots are allocated at events hosted by different countries. A first observation is that events and judging slots are allocated roughly in

proportion to countries' shares of competitor performances. Notable exceptions are Ukraine and other Asia-Pacific countries (e.g., Chinese Taipei and Australia), which host fewer events. A second observation is that countries are allocated a higher share of the judging slots in events they host and, to a lesser extent, in events located in the same region. Regression analysis (omitted for space reasons) confirms that the host-country and same-region effects on judging slot allocations are statistically significant, and also reveals that judging slot allocations are negatively correlated within seasons, suggesting that there is an effort to target a certain allocation of slots across countries.

Tables 4 and 5 present estimates of bias using equation (1) for the 6.0 and code of points judging systems, respectively. Results are shown for technical merit and artistic impression scores from the pre-scandal and anonymous-6.0 samples, as well as for the major components of scores from the code of points sample. In each regression, the unit of observation is a performance. The first specification includes only the compatriot-judge indicator variable, and the second specification includes controls for home-country and home-region. Standard errors in all specifications allow for clustering of residuals within both competitor-country and competition, using the procedure outlined in Petersen (2009).[18]

Two additional specifications are reported for the code of points system. Since the sample period for the code of points analysis covers many seasons, a third specification replaces competitor fixed effects with competitor*season fixed effects. A fourth specification replaces competitor fixed effects with competitor*event fixed effects. This last specification identifies judging biases using only events with different judging panels for different rounds in the same competition.[19]

[18] Clustering for competitor-country alone instead of both competitor-country and competition produces standard errors that are only 2-3 percent smaller for most of the specifications in Table 5. Clustering for competition alone has a much bigger impact, producing standard errors that are about 25-30 percent smaller.

[19] About 24 percent of the performances in the code of points sample occurred in competitions with different judging panels for different rounds of the same competition. This includes most competitions at the European, World, and World Junior Championships, the Four Continents championship in 2005, and Grand Prix events in

We can make several observations from the tables. First, the results are consistent across specifications. Adding the home-country and home-region controls reduces estimates of bias, but only by approximately 10 percent. Switching to competitor*season fixed effects reduces the precision of the estimates, and increases the magnitude of estimated biases slightly for most outcomes. Switching to competitor*event fixed effects reduces the precision of the estimates, but the estimates for most component scores remain statistically significant, and the changes are not statistically significant for any outcome.

A second observation is that we find evidence of a compatriot-judge effect for technical merit and artistic impression in the 6.0 system and for grades of execution and for all five program component scores in the code of points system. We do not find evidence of a compatriot-judge effect for base values, which are assigned based on a formula, and deductions, which are assessed for obvious faults such as falls (any fault requiring a judgment, such as a takeoff that may have been from the wrong edge, is referred to the judges and reflects separately by each judge in their grade of execution marks). In other words, compatriot-judge effects are evident only in subjective scores given by individual judges, and not in objective measures of a program's difficulty and execution. This suggests that compatriot-judge effects are indeed due to judging biases and not to competitors not altering their skating when they have a compatriot on the judging panel.[20]

In results omitted for space reasons, I also tested specifications that included a variable for compatriot in the roles of referee (who oversees the judging panel) and technical controller (who leads the technical panel that assigns base values and identifies deductions). These positions were analyzed

---

2003-4. Pairs skating was less likely than the other disciplines to have different panels (only 10 percent of performances were in such events).

[20] As an additional falsification test, I checked whether scores were increasing in the number of judges with whom a competitor (or, for pairs, either of the competitors) shared a first letter of either a first or last name. I found no evidence of this, so long as same nationality was controlled for (same nationality and same first letter are positively correlated for both first and last names). Estimates were not only statistically insignificant but also reasonably precise, with standard errors that were about half of those for the compatriot judge effect.

both because they are arguably the two non-judge positions with the most authority and they are also the two positions that are usually held by officials who also serve as judges, which is important since it facilitates determining their home countries.[21] I found no evidence that compatriots in these positions had a statistically significant effect on any components of scores.[22] In particular, there was no evidence that the nationality of the technical coordinator affected base values or deductions, possibility because those determinations were more objective, or possibly because the technical panel involves multiple individuals and this may limit discretion.

One might view technical merit as more objective than artistic impression, grades of execution as more objective than program components, and program components as being listed in approximate order of objectivity. For instance, the first two program components, skating skills and transitions, were components of technical merit under the 6.0 system, while the last two, choreography and interpretation, were components of artistic impression. Under this view, we consistently find larger judging biases (in points and in within-round and within-performance standard deviations) for more subjective scores.

Have judging biases increased or decreased under anonymity? Table 6 summarizes the estimates of judging bias from the different samples and compares them to within-round and within-performance standard deviations of scores. Comparing the two 6.0-system samples, we find judging biases are roughly 20 percent larger (in points and both varieties of standard deviations) when judges are anonymous, despite the fact the increase in average judging panel size (from 8.4 to 11.1; see Table

[21] Home countries of judges are generally disclosed for Grand Prix and Junior Grand Prix events, but judges are listed as representing the ISU in major championships, and all non-judge officials are always listed as representing the ISU. Fortunately, judges in the major championships and the referees and technical controllers almost always appear elsewhere in my sample with their nationality disclosed. In the few cases where this was not the case, I was able to identify their nationality from the pre-scandal data or from the countries represented by judges in their careers as competitors. In contrast, I was usually not able to identify the nationality of other officials (e.g. technical specialists, assistant technical specialists, data operators, video operators).

[22] In some less important events, the referee will also play the role of a judge. In these cases, competitors from this judge's country were coded as having both a compatriot judge and a compatriot referee, and the compatriot referee coefficient captured the incremental effect of having a judge who also served in the referee role.

1) should have diluted the direct effect of a compatriot judge.[23] Larger panels should also have reduced incentives for judging biases. If judges' optimal bias reflects a tradeoff between influencing results and not appearing out of line with other judges, then larger judging panels should reduce biases for two reasons: they reduce the influence any one judge has on the results, and they increase the number of peers with which one can be compared.[24] The fact that judging bias increased despite larger judging panels suggests that judge anonymity did not have its hoped for effect.

In addition, Table 4 reveals that the home-country coefficient increased sharply after judges were anonymized.[25] This coefficient cannot be regarded as a pure measure of judging bias, as it would also reflect a tendency for competitors to perform better at home. But assuming that the home-ice effect on true performance remained constant, then the change in this coefficient might reflect changes in judging biases. Unfortunately, due to judge anonymity, I cannot determine whether this change reflects more nationalism among host-country judges or a greater pro-host bias among other judges.

Comparing the code of points and pre-scandal samples is less straightforward. If we compare technical merit scores with grades of execution and artistic impression scores with program components and normalize using within-round standard deviations, we find that judging biases are larger in the more recent sample for technical merit/execution and about the same for artistic impression/components. There is only slightly more judging bias in total scores under the code of points system, however, since about half of the total variance in total score is now determined by a program's base value and mandatory deductions, over which individual judges do not have discretion.

---

[23] Adjusting results for panel size, as suggested by a referee, is not straightforward, as the compatriot-judge effect arises from both a nationalistic bias from the compatriot judge and better scores from the judges. In Zitzewitz (2006), I report that the former benefit is about 3 times as larger as the latter in the pre-scandal sample. If these effects are constant, then their effect on a 8-judge average would be about 6 percent larger than on a 11 judge panel [(7+1*3)/8]/[(10+1*3)/11] = 110/104.

[24] A referee points out a potential offsetting effect. When scores are aggregated using a trimmed mean, more judges means that the second highest score, which is the cutoff at which a positive outlier score stops being influential, is likely to be higher. This may cause larger judging panels to increase incentives for bias for scores in some ranges, but I was unable to construct an example where this effect outweighed the reduction in incentives from the larger number of scores being averaged.

[25] I thank a referee for drawing my attention to this.

Judging biases have in contrast decreased slightly under the code of points system when normalized by within-performance standard deviations. This reflects the fact that within-performance standard deviations are relatively larger under the code of points system. For evaluations of the "fairness" of a judging system in the sense of judging panel composition not affecting competitors' placements, normalization by the within-round standard deviation is more appropriate. Normalization by within-performance standard deviation is useful because it gives a measure of how obvious a judging bias is likely to be to those privy to individual judges' scores. The greater variation in judges' scoring of the same performance under the code of points system may make a given size judging bias less salient to the judges, as well as to any officials charged with monitoring for bias.

Table 7 presents estimates of judging biases for subsamples of the code of points sample. Biases are greater for events with smaller judging panels. Biases in both the components and grades of execution are largest in ice dancing, second largest in ladies' events, and smallest in men's and pairs. Larger biases in ice dancing are consistent with ice dancing involving less emphasis on jumps, leaving a larger role for subjective performance evaluation. Biases are larger in major championships and grand prix events and smaller in junior events. There is no clear trend when comparing across seasons -- biases in components scores have increased slightly and biases in GOE scores have decreased, leaving the total bias essentially unchanged. In unreported results, biases are not systematically larger for skaters for top performers (based on their performance in the most recent competition) nor for those randomly assigned a later starting position in the competition's initial round.[26]

Given the differences in the proportion of each sample accounted for by different levels of competition, a referee suggested re-weighting the code of points sample to match the mix of events in the pre-scandal sample. Doing so puts more weight on the major events that exhibit more judging bias. For example, the compatriot-judge effect on total scores rises from 0.612 (SE 0.226) to 0.768 (SE 0.243).

[26] Bruine de Bruin (2006) finds evidence that skaters who are randomly assigned later starting positions do earn higher scores.

As all pre-scandal events used fewer than 12 judges, an analogous reweighting based on number of judges yields the much higher estimate reported in the first line of Table 7 (1.488; SE 0.646). Thus the conclusion that biases have worsened since the 2002 scandal would be even stronger if differences in the mix of events across samples was controlled for.[27]

V. Indirect evidence on nationalistic bias and vote trading

One might expect anonymity to affect the compatriot-judge effect in two offsetting ways. Anonymity may have its intended effect of reducing vote trading by making defecting on collusive agreements harder to detect. At the same time, anonymity may frustrate external monitoring of individual judges' biases. Although the ISU states that it also monitors judges, the loss of external monitoring may lead to more nationalistic bias.

The estimates in Zitzewitz (2006) imply that 25-30 percent of the total impact of a same-country judge on mean scores came from the judge in question, with the other 70-75 percent coming from the impact of a same-country judge on the other judges' scores. There was also some evidence that this effect was not uniform, with judges voting in blocs that resembled those alleged to have influenced the 2002 Olympic Pairs competition.[28] One might have expected the 25-30 percent share to have grown under anonymity, for the reasons discussed above.

Anonymity unfortunately frustrates the straightforward approach to separating nationalistic bias and vote trading taken in my earlier work, so I turn to less direct evidence. If nationalistic bias has increased in importance relative to vote trading, we might expect to see a single positive outlier score when a compatriot is on the panel. Table 8 presents versions of specification 3 from Table 5 that

[27] In contrast, the proportion of each sample accounted for by Ice Dancing, which exhibits larger biases, was much more constant.

[28] China, France, Poland, Russia, Ukraine were alleged to have voted in one bloc, with Canada, Germany Italy, and the U.S. in the other bloc.

substitute the maximum-to-median spread as a dependent variable. Effects on the inter-quartile range are provided as a comparison.[29]

The results suggest that max-median and inter-quartile spreads were no larger for panels with compatriot judges in the pre-scandal and anonymous-6.0 samples. In the code of points sample, both spreads are larger with a compatriot judge: for the sum of grades of execution and components scores (the two components of total score on which judges can differ), the max-median spread is 0.22 points larger, compared with a mean spread of 6.41 points reported in Table 2. The inter-quartile spread is 0.14 points larger, compared with a mean of 4.56. Thus the two spreads increase roughly in proportion, suggesting a general increase in dispersion when a same-country judge is on the panel, rather than the single outlier one might expect if the effect were purely due to a nationalistic bias.

The analysis of spreads does not suggest a significant shift in the composition of the compatriot judge effect. I now turn to analyses that test for the continued presence of both components. Specification 1 in Table 9 includes a variable that counts the number of compatriot judges on the judging panel in other disciplines at the same event. The goal is to test for the form of vote trading that was alleged at the 2002 Olympics, where a Russian pair was said to have been aided by a trade for a Russian judge's vote in Ice Dancing. The results suggest evidence of small biases in a direction consistent with vote trading. For example, whereas a compatriot on the judging panel in one's own discipline is estimated to raise the sum of GOE and components scores by 0.390, a compatriot on another discipline's judging panel is estimated to raise the same aggregate by 0.094. This suggests that vote trading has continued under the code of point system, but also that a compatriot on one's own discipline's panel is of much greater benefit to a competitor. This last result could either reflect direct

[29] I could have alternatively used the ratio of the max-to-median and inter-quartile spreads for each performance as the dependent variable. Many performances have inter-quartile ranges that are quite small, and so the results from this approach would be sensitive to these outliers.

nationalistic bias being a larger component of the total compatriot-judge effect than in the pre-scandal sample or it could reflect vote trading is easier to implement within a given discipline.

Additional specifications in Table 9 test whether compatriot-judge effects are larger for certain countries. In Zitzewitz (2006), I found that larger nationalistic biases from judges from countries that were regarded as less transparent (as reflected in the 2001 version of Transparency International's *Corruption Perceptions Index*). In specifications 2 and 3, I find that the compatriot-judge effect in the code of point sample (2003-9) is larger for countries that had larger nationalistic biases in 2001-2 and for those regarded as less transparent.

Specifications 4 and 5 test for the persistence of the bloc judging identified in my earlier work. In Zitzewitz (2006), I tested all possible permutations of a two-bloc judging model and found that the pattern of cross-country biases among the top 10 countries (in terms of competitor participation) was best explained by a model with a western bloc that included Canada, Germany, Italy, and the United States, an eastern bloc that included France, Poland, Russia, and Ukraine, and two unaffiliated countries, China and Japan. These results were interesting in part because of the fact they were consistent with the alleged voting blocs in the 2002 Pairs competition, despite the fact that I omitted that competition when estimating the model. They were also consistent with Seltzer and Glass (1991) and Sala, Scott and Spriggs (2007), who found evidence of bloc judging along cold war lines (albeit with France tilting slightly to the East) in earlier Olympics.

Given judge anonymity, I lack the necessary statistical power to determine whether these specific blocs still best explain the data, but I can test whether the pattern of cross-bloc biases in the code of points sample is consistent with my earlier results. In Specification 4, I find that competitors from Western and Eastern bloc countries benefit from compatriot-judge effects that are roughly the same size as those of other countries. In Specification 5, I find that the only statistically significant cross-

bloc effect is a negative effect of western-bloc judges on eastern-bloc competitors. In both respects, these results are consistent with what is reported in Table 8 of Zitzewitz (2006).

Taken together, the results in Table 8 and 9 suggest a persistence of both nationalistic biases and vote trading. The analysis of spreads in Table 8 does not suggest a significant difference in the composition of the compatriot-judge effect in the pre-scandal to the code of points samples. Specification 1 of Table 9 provides evidence that competitors benefit from having compatriot judges in other disciplines, which may be due to vote trading. Specifications 2 and 3 suggest that judge countries that were the most nationalistically biased in the past provide their competitors with greater compatriot-judge effects in the current data, hinting at a persistence of nationalistic bias. As acknowledged in this section's title, this evidence is all relatively indirect; more direct evidence would be possible if the mapping of scores to judges is eventually disclosed.

VI. Alternative explanations

All of the tests for compatriot-judge effects in this paper test whether results are correlated with something I argue they should not be -- the composition of the judging panel. In interpreting the correlations I find as evidence of judging bias, I am assuming that, once home-country, home-region, and competitor and round fixed effects are controlled for, there is no reason why a competitors' true performance quality vary with the composition of the judging panel.

Support for this assumption is found in the fact that those components of competitors' scores that are objectively determined (base value and mandatory deductions) do not vary with judging panel composition, suggesting that competitors skate no differently when they have compatriots on the panel. Further support is found in the fact that controlling for the most obvious observable variables that might be correlated with both true performance and panel composition has very small effects on the bias estimates.

Controlling for home country and home region effects reduces estimated biases as one might expect, but estimates decline by only approximately 10 percent. Event location is clearly a leading candidate for a variable that might affect both competitor performance and panel composition, since host countries select the judge countries and since travel-related fatigue might affect competitors' true performance. Given that controlling for this omitted variable has such a limited effect on the coefficient of interest, it increases our confidence that the results will be robust to other potential omitted variables.

Another potential issue is that the selection of competitors into events may depend on the panel composition. In unreported results, I find that competitors are slightly more likely to participate in events that are in their home country and that have compatriot judges. These effects are larger among the half of competitors with the lowest estimated fixed effects, suggesting that marginal competitors are more likely to participate in events in nearby locations and with compatriot judges.

While my approach controls for competitor fixed effects, a version of this selection effect might operate within competitors. Competitors might participate in all events when they expect to perform well, but only in events that are close to home or have friendly judges when they expect to perform poorly. This would create a *negative* within-competitor correlation between true performance both home-ice and compatriot judges, downwardly biasing estimates of both compatriot-judge and home-country effects. If such an effect were important, we should presumably see it reflected in our results for base values. Small within-competitor selection effects may explain why estimated compatriot-judge and home-country effects both rise slightly when competitor fixed effects are replaced with competitor*season fixed effects. So long as any within competitor*season selection effects are in the same direction, they would bias us (slightly) against finding compatriot-judge biases.

One important limitation of my analysis is that while I can determine that evaluations of a skater from country A are better when A is represented on the panel than when it is not, I cannot determine

which evaluations are more correct. A compatriot-judge effect could arise from the judge creating an unfair bias in favor of her compatriots, or from the compatriot-judge negating what would otherwise be an unfair bias against her compatriots. Likewise, when we find that Eastern-bloc competitors receive lower scores when there are more Western-bloc judges, this could arise either from the Western judges biasing against Eastern competitors, or from all other judges biasing in their favor.

This last point raises the issue of tastes. Tastes for different styles of figure skating that were correlated within countries (or western and eastern blocs) could explain some of the results in this paper, as well as some of the results in earlier work (Bassett and Persky, 1991; Campbell and Galbraith, 1996; Zitzewitz, 2002 and 2006; Sala, Scott, and Spriggs, 2007). Other results are more difficult for tastes to explain. For example, in Zitzewitz (2006), I find that the scores given to a competitor from country A by a judge from country B varies depending on whether A was represented on the judging panel. As discussed above, these cross-country effects accounted for 70-75 percent of the overall compatriot-judge effect in the 2001-2 sample, and there was little evidence in Section V that this proportion has changed significantly. Likewise, in this paper, I find that the scores given to competitors from country A vary depending whether A is represented on the judging panel in other disciplines.

Furthermore, I find that while bias estimates are larger for the more subjective dimensions of performance (Artistic Impression, Choreography, Interpretation), on which there is arguably more scope for national styles and tastes, biases are nearly as large for grades of execution, which reflect very technical issues like whether a skater used the correct skate edge. While tastes may be correlated within countries, they are unlikely to explain a significant portion of the compatriot-judge effect.

VII. Conclusion

I have shown in this paper that figure skaters benefit from a compatriot on their judging panel, that this benefit likely reflects a combination of nationalistic bias and vote trading, and that this benefit has risen

slightly (albeit not statistically significantly) over time. The increase in the combination of bias and vote trading was despite a reform that was purportedly intended to reduce it. A key component of that reform was eliminating transparency into which judge gave which score. Eliminating transparency was designed to make it harder for parties to collusive agreements to monitor judges, but this came at the cost of making monitoring by outsiders harder as well. The net of these two effects is theoretically ambiguous, motivating the empirical analysis in this paper.

When anonymity was introduced, the ISU sought to allay concerns by promising to conduct extensive internal monitoring of judging bias. The results of this paper suggest that this internal monitoring was not as effective as needed for anonymity to have a net negative effect on bias. One of the great advantages of the external monitoring facilitated by transparency is that there is free entry into the role of monitor. There is therefore likely to be competition among outside monitors that helps create incentives for truthful revelation. If an internal monitor is a monopolist, there is more scope for capture by interested parties or for the monitors' personal tastes to affect outcomes.

One can thus view the ISU's anonymity reform as a well-intentioned attempt to reduce corruption that failed due to insufficiently effective internal monitoring. A less optimistic view is that the ISU's goal was to reduce the perception of corruption rather than actual corruption. Perceptions of corruption need not be fully accurate, and if limited attention leads spectators to underestimate corruption in the absence of hard evidence thereof, then reducing transparency can be an end in itself. Some of the actions of the ISU after the 2002 judging scandal can only be rationalized as attempts to reduce the perception of corruption by limiting outside monitoring. Examples include obscuring which judges gave which score for events where this had been previously disclosed, the ongoing practice of not disclosing judges' nationalities at major competitions, and the very recent practice of shuffling judges scores mentioned by Emerson and Arnold (2011). These changes are unlikely to frustrate insiders attempting to enforce a collusive agreement, but do raise the cost of outside analyses like this one.

These issues are all present in many settings outside sports. One important point often lost in discussions of the costs and benefits of transparency is that transparency is multi-dimensional. In particular, policies creating transparency differ in the costs they impose on those seeking to access data. When access costs are high (e.g., suing under the Freedom of Information Act), transparency is likely to disproportionately inform those willing to pay the access costs, who are more likely to be insiders. Low-access-cost policies (i.e., posting data on the internet in easily analyzed formats) in turn are likely to differentially inform larger numbers of outside analysts, and thus may create the competition among analysts conducive to truthful revelation.

This suggests an alternative to the problems that organizations sometimes seek to solve by reducing transparency: more transparency, but of a different form. For example, if one is concerned that disclosing mutual fund proxy votes will facilitate management's influence of these votes through the offer of special favors, two possible solutions would be to: 1) disclose the data in a form most likely to be useful to outside monitors, and 2) disclose data capturing the most obvious forms of favors. If one is concerned that roll call votes will facilitate monitoring by special interests, the two analogous solutions would be to: 1) post information on roll call votes in a form easily accessible by votes and the media and 2) improve disclosure on the earmarks, and other channels through which special interests benefit.

Returning to figure skating, the results of this analysis suggest that a return to 2002 levels of transparency would reduce judging biases. The sport might improve upon 2002 policy though by providing disclosure in a different form. Collecting the data for this project required a substantial amount of time spent parsing of HTML and PDF files. A policy change as seemingly trivial as making this data available as a preassembled dataset would lower the costs of outside monitoring, while having no effect on the enforcement of collusive schemes. Such a change would reflect a sharply different overall

approach, away from what could be viewed as an (ultimately unsuccessful) attempt to obscure a problem, and towards harnessing willing outsiders to help solve it.

References

Arnold, Douglas. 1983. "Can Inattentive Citizens Control Their Elected Representatives?" in Lawrence C. Dodd and Bruce Oppenheimer (eds.), *Congress Reconsidered*, 5th ed. (Washington: CQ Press, 1993), pp. 401-416.

Bassett, Gilbert and Joseph Persky. 1994. "Rating Skating," *Journal of the American Statistical Association* 89(427) 1075-1079.

Bruine de Bruin, Wandi. 2006. "Save the Last Dance II: Unwanted Serial Position Effects in Figure Skating Judgments," *Acta Psychologica* 123, 299-311.

Campbell, Bryan and John Galbraith. 1996. "Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments," *The Statistician* 45(4), 521-526.

Chen, Jowei. 2007. "What Does Baseball Teach Us About Reducing Racial Discrimination? Evidence from Two Natural Experiments," unpublished.

Davis, Gerald and Han Kim. 2007. "Business ties and proxy voting by mutual funds."*Journal of Financial Economics* 85, 552-57.

Dohmen, Thomas. 2005. "Social pressure influences decisions of individuals: Evidence from the behavior of football referees," *IZA Discussion Paper* No. 1595.

Duggan, Mark and Steven Levitt. 2002. "Winning Isn't Everything: Corruption in Sumo Wrestling," *American Economic Review* 92, 1594-1605.

Emerson, John. 2007. "Chance, On and Off the Ice," *Chance* 20(2).

Emerson, John, Miki Seltzer, and David Lin. 2009. "Assessing Judging Bias: An Example from the 2000 Olympic Games," *The American Statistician* 63(2), 124-131.

Emerson, John and Taylor Arnold. 2011. "Statistical Sleuthing by Leveraging Human Nature: A Study of Olympic Figure Skating," *The American Statistician* 65(3), 143-148.

Fenwick, Ian and Sangit Chatterjee. 1981. "Perception, Preference and Patriotism: An Exploratory Analysis of the 1980 Winter Olympics," *The American Statistician* 35(3), 170-173.

Garicano, Luis, Palacios-Huerta, Ignacio and Canice Prendergast. 2005. "Favoritism under social pressure," *Review of Economics and Statistics* 87, 208-216.

Gordon, Stephen and Michael Truchon. 2008. "Social Change, Optimal Inference, and Figure Skating," *Social Choice and Welfare* 30, 265-284.

Green, Edward and Robert Porter. 1984. "Non-cooperative Collusion Under Imperfect Price Information," *Econometrica* 52(1), 87-100.

Lee, Jungmin. 2008. "Outlier Aversion in Subjective Evaluation," *Journal of Sports Economics* 9(2), 141-159.

Levitt, Steven. 2002. "Testing the economic model of crime: the National Hockey League's two referee experiment," *Berkeley Electronic Journals in Economic Analysis and Policy*, 1, 1-19.

Loosemore, Sandra. 2002. "Editorial Comment on the Proposed ISU Scoring Changes," available at http://www.frogsonice.com/skateweb/articles/editorial-isu-changes.shtml (last accessed August 12, 2010).

Moul, Charles and John V.C. Nye. 2009. "Did the Soviets Collude? A Statistical Analysis of Championship Chess, 1940-1978," *Journal of Economic Behavior and Organization* 70(1-2), 10-21.

Parsons, Christopher, Johan Sulaeman, Michael Yates, and Daniel Hamermesh. 2011. "Strike Three: Umpires' Demand for Discrimination," *American Economic Review* 101(4), 1410-35.

Petersen, Mitchell. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *Review of Financial Studies* 22, 435-480.

Prendergast, Candice. 1993. "A Theory of 'Yes Men'," *American Economic Review* 83(4), 757-770.

Price, Joseph, Marc Remer, and Daniel Stone. 2012. "Sub-Perfect Game: Profitable Biases of NBA Referees," *Journal of Economics and Management Strategy* 21(1), forthcoming.

Price, Joseph and Justin Wolfers. 2010. "Racial Discrimination Among NBA Referees," Quarterly Journal of Economics 125(4), 1859-1887.

Sala, Brian, Scott, John, and James Spriggs. 2007. "The Cold War on Ice: Constructivism and the Politics of Olympic Skating Judging," *Perspectives on Politics* 5(1), 17-29.

Seltzer, Richard and Wayne Glass. 1991. "International Politics and Judging in Olympic Skating Events: 1968-1988," *Journal of Sports Behavior* 14, 189-200.

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

Stigler, George. 1964. "A Theory of Oligopoly," *Journal of Political Economy* 72(1), 44-61.

Wolfers, Justin. 2006. "Point Shaving: Corruption in NCAA Basketball," *American Economic Review* 96(2), 279-283.

Wu, Samuel and Mark Yang. 2004. "Evaluation of the Current Decision Rule in Figure Skating and Possible Improvements," *The American Statistician* 58(1), 46-54.

Zitzewitz, Eric. 2002. "Nationalistic Bias in Winter Sports Judging and its Lessons for Organizational Decision Making," Stanford GSB Working Paper No. 1796.

Zitzewitz, Eric. 2006. "Nationalistic Bias in Winter Sports Judging and its Lessons for Organizational Decision Making," *Journal of Economics and Management Strategy* 15(1), 67-99.

Table 1. Sample Size

Seasons	Pre-scandal 2000-1 (part) and 2001-2	Anonymous 6.0 2002-3 and 2003-4*	Code of Points 2003-4* to 2008-9
Unique events	16	23	107
Major championship (Olympics, European, Four Conts., World)	8	6	16
Grand Prix	4	7	42
Junior (Junior Grand Prix and World Junior Championship)	4	10	49
Competitions (Men, Ladies, Pairs, Dancing)	61	92	416
Rounds	181	234	950
Performances	2,976	3,678	15,159
Evaluations	25,068	40,844	154,964
Judge evaluations per performance	8.4	11.1	10.2
Unique competitors	584	627	1,486
Performances per competitor	5.1	5.9	10.2
Competitors per round	16.4	15.7	16.0
Share of competitors with compatriot judge	0.528	0.688	0.684
Share of judges that are compatriots	0.063	0.062	0.067

* Grand Prix events in the 2003-4 season used the Code of Points judging system and are included in that sample.

Table 2. Summary statistics

Average trimmed mean is calculated by taking the trimmed mean (i.e. the average of all but the highest and lowest scores) of judges' scores of a given performance and then taking the average across all performances. The within-round standard deviation is the standard deviation of trimmed means across different competitors in a given round of a given competition. The within performance standard deviation is the standard deviation of the judges' scores given to a specific performance. Note that summary statistics for TM+AI are not available for the Anonymous 6.0 sample since the pairing of TM and AI scores given by the same judge was not disclosed.

	Mean	Standard deviations		Other within performance spreads		
	Trimmed mean	Within round	Within performance	Max-Median	75-25 spread	Ratio
Panel A. Pre-scandal sample						
Total (TM+AI)	9.21	1.23	0.35	0.47	0.42	1.10
Technical merit (TM)	4.49	0.66	0.20	0.26	0.22	1.16
Artistic impression (AI)	4.72	0.58	0.19	0.24	0.23	1.05
Panel B. Anonymous 6.0 sample						
Technical merit (TM)	4.43	0.66	0.21	0.31	0.26	1.21
Artistic impression (AI)	4.68	0.57	0.20	0.28	0.25	1.16
Panel C. Code of points sample						
Total Segment Score (TES + TPCS - D)	58.99	12.01	4.45	6.41	4.56	1.41
Total Element Score (TES = BV + GOE)	30.56	6.77	NA	NA	NA	1.75
Base Value (BV)	31.58	5.49	NA	NA	NA	NA
Grade of Execution (GOE)	-1.02	2.74	NA	NA	NA	1.75
Total Program Component Score (weighted sum)	29.66	5.95	3.31	3.34	2.70	1.24
Skating Skills	5.16	0.96	0.64	0.59	0.48	1.24
Transition / Linking Footwork	4.85	0.97	0.65	0.65	0.53	1.24
Performance / Execution	5.04	0.98	0.65	0.64	0.52	1.24
Choreography / Composition	5.04	0.96	0.66	0.65	0.52	1.24
Interpretation / Timing	5.04	0.98	0.67	0.66	0.54	1.23
Unweighted sum	24.62	4.73	2.58	2.86	2.31	1.24
Deductions (D)	0.46	0.67	NA	NA	NA	NA
Components score (unweighted) + GOE	23.60	8.97	3.91	6.16	4.24	1.45

Table 3. Events hosted, competitors, and judging slots by country

The top panel of this table reports major countries' share of events hosted, judging slots, and competitor performances. Major countries are defined as the top 10 countries in terms of competitor performances, judging slots, and events hosted during the sample period. The bottom panel reports the allocation of judging slots for events hosted by a country. The bottom panel compares the allocation of judging slots for events hosted by different major countries.

	Host country												
	Americas			Asia-Pacific			Europe/Middle East/Africa						
	USA*	CAN*	Other	JPN*	CHN*	Other	RUS*	FRA*	UKR	ITA	GER	POL	Other
Share of:													
Performances	12%	10%	1%	6%	5%	6%	10%	6%	4%	4%	4%	2%	31%
Events hosted	11%	10%	2%	8%	7%	1%	7%	9%	1%	4%	4%	3%	34%
Judging slots	9%	7%	0.4%	5%	4%	6%	8%	7%	5%	4%	6%	4%	37%
Average number of judging slots allocated to country listed below													
Americas													
USA	**0.91**	0.89	1.00	0.97	0.91	1.00	0.82	0.83	0.56	0.73	0.58	0.52	0.72
CAN	0.88	**0.91**	1.00	0.90	0.89	0.78	0.81	0.85	0.56	0.73	0.66	0.33	0.66
Other Americas	0.05	0.02	0.24	0.02	0.00	0.67	0.00	0.02	0.00	0.00	0.08	0.07	0.04
Asia-Pacific													
JPN	0.60	0.69	0.60	**0.92**	0.64	0.44	0.60	0.37	0.22	0.61	0.44	0.41	0.35
CHN	0.42	0.46	0.28	0.76	**0.82**	0.00	0.52	0.45	0.00	0.24	0.44	0.22	0.18
Other Asia-Pacific	0.76	0.62	0.28	0.97	1.09	1.67	0.38	0.32	0.56	0.48	0.54	0.26	0.48
Europe/Middle East/Africa													
RUS	0.78	0.81	0.56	0.93	0.92	0.78	**0.90**	0.92	0.56	0.67	0.56	0.78	0.71
FRA	0.60	0.57	0.40	0.47	0.39	1.00	0.77	**0.95**	0.56	0.76	0.56	0.78	0.51
UKR	0.33	0.41	0.20	0.45	0.36	0.78	0.54	0.54	**0.78**	0.58	0.58	0.78	0.56
ITA	0.37	0.44	0.36	0.33	0.26	0.00	0.45	0.46	0.44	**0.70**	0.44	0.56	0.50
GER	0.58	0.72	0.44	0.56	0.49	0.78	0.48	0.64	0.78	0.52	**0.78**	0.52	0.55
POL	0.34	0.23	0.00	0.27	0.36	0.00	0.42	0.26	0.56	0.18	0.50	**0.70**	0.42
Other Europe	3.35	3.56	2.08	2.25	2.75	1.78	3.44	3.28	4.22	4.85	5.10	4.63	4.52
Allocations of judging slots by region:													
Americas	20%	19%	32%	21%	20%	33%	17%	18%	12%	14%	12%	9%	15%
Asia-Pacific	19%	18%	17%	30%	29%	29%	15%	12%	8%	13%	13%	9%	10%
Europe/Middle East/Africa	69%	70%	58%	60%	63%	70%	72%	74%	86%	78%	80%	86%	80%

* These six countries each host a Grand Prix event each season.

Table 4. Compatriot judge effects in the 6.0 scoring system

This table presents estimates of equation 1 for trimmed means of technical merit (TM), artistic impression (AI), and their sum for the pre-scandal and anonymous-6.0 samples. All regressions include fixed effects for rounds (event* competition* round combinations) and competitors. Standard errors (in parentheses) adjust for clustering within competitor country and competition.

	Pre-scandal sample			Anonymous 6.0 sample		
	TM+AI	TM	AI	TM+AI	TM	AI
Specification 1. Baseline						
Compatriot judge	0.051	0.017	0.033**	0.063*	0.030	0.032**
	(0.044)	(0.029)	(0.016)	(0.034)	(0.021)	(0.015)
Specification 2. Controls for home country and region						
Compatriot judge	0.053	0.019	0.033**	0.064*	0.031	0.033**
	(0.043)	(0.023)	(0.016)	(0.035)	(0.021)	(0.015)
Home country	-0.006	-0.024	0.018	0.124**	0.059**	0.066***
	(0.083)	(0.051)	(0.032)	(0.050)	(0.029)	(0.022)
Home region	-0.058	-0.031	-0.027	0.003	0.005	-0.002
	(0.055)	(0.029)	(0.029)	(0.042)	(0.025)	(0.019)
Observations	2,976	2,976	2,976	3,678	3,678	3,678
R-squared (Specification 1)	0.908	0.855	0.927	0.898	0.84	0.921

Significance at the 10, 5, and 1 percent level is indicated with *, **, and ***.

Table 5. Compatriot judge effects in the code of points system

This table presents estimates of equation 1 for trimmed means of different score elements for the code of points sample. All regressions include fixed effects for rounds (event*competition*round combinations). Specifications 1 and 2 include fixed effects for competitors, specification 3 includes fixed effects for competitor*season combinations, and specification 4 includes fixed effect for competitor*event combinations. Specification 3 also includes the host country and home region controls. Standard errors (in parentheses) adjust for clustering within competitor country and competition.

		Total Element Score			Program Components Score							
	Total score	Total	Base Value	GOE	TCPS	Skating skills	Transitions	Performance	Choreography	Interpretation	Sum	Deductions
Specification 1. Baseline												
Compatriot judge	0.370*	0.033	-0.116	0.149**	0.269***	0.047***	0.059***	0.056***	0.061***	0.061***	0.282***	0.0232
	(0.193)	(0.125)	(0.109)	(0.059)	(0.081)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)	(0.059)	(0.018)
Specification 2. Controls for home country/region												
Compatriot judge	0.307	-0.004	-0.133	0.128**	0.236***	0.042***	0.055***	0.050***	0.056***	0.054***	0.255***	0.0218
	(0.194)	(0.126)	(0.109)	(0.059)	(0.082)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)	(0.060)	(0.044)
Home country	0.445*	0.212	0.0473	0.165**	0.295***	0.052***	0.046***	0.063***	0.051***	0.0695***	0.273***	0.0515
	(0.259)	(0.168)	(0.138)	(0.083)	(0.104)	(0.016)	(0.016)	(0.017)	(0.016)	(0.018)	(0.079)	(0.044)
Home region	0.485**	0.349**	0.214*	0.136*	0.191**	0.0179	0.0219	0.0257*	0.0293**	0.0287*	0.119*	-0.039
	(0.226)	(0.149)	(0.116)	(0.075)	(0.089)	(0.013)	(0.014)	(0.015)	(0.014)	(0.015)	(0.068)	(0.044)
Specification 3. Competitor*season fixed effects												
Compatriot judge	0.612***	0.187	0.023	0.164**	0.320***	0.049***	0.063***	0.057***	0.060***	0.069***	0.295***	-0.006
	(0.215)	(0.140)	(0.123)	(0.070)	(0.089)	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)	(0.062)	(0.020)
Home country	0.538**	0.185	-0.0420	0.227**	0.385***	0.0574***	0.0519***	0.0728***	0.0601***	0.0830***	0.316***	0.0454
	(0.260)	(0.179)	(0.153)	(0.0918)	(0.0984)	(0.0150)	(0.0153)	(0.0169)	(0.0153)	(0.0169)	(0.0754)	(0.0283)
Home region	0.483**	0.463***	0.375***	0.0884	0.101	0.000241	0.00190	0.00823	0.00646	0.00331	0.0183	-0.0268
	(0.231)	(0.163)	(0.132)	(0.0858)	(0.0848)	(0.0127)	(0.0131)	(0.0146)	(0.0131)	(0.0148)	(0.0649)	(0.0224)
Specification 4. Skater*round effects												
Compatriot judge	0.434	-0.908	-1.15	0.0499	0.261	0.029*	0.051***	0.039**	0.028	0.0555**	0.217*	0.0308
	(0.600)	(1.032)	(1.034)	(0.176)	(0.269)	(0.016)	(0.017)	(0.019)	(0.027)	(0.023)	(0.113)	(0.049)
Observations	15,159	15,159	15,159	15,159	15,159	15,159	15,159	15,159	15,159	13,649	15,159	15,159
R-squared (Specification 1)	0.95	0.916	0.933	0.623	0.965	0.949	0.943	0.939	0.944	0.94	0.945	0.372

Significance at the 10, 5, and 1 percent level is indicated with *, **, and ***.

Table 6. Comparing the compatriot judge effect across samples

Bias estimates are from specification 2 in Tables 4 and specification 3 in Table 5; standard deviations are from Table 2.

		In standard deviations	
	Bias estimate	Within round	Within performance
Pre-scandal			
Total (TM+AI)	0.053	0.043	0.151
Technical merit (TM)	0.019	0.028	0.096
Artistic Impression (AI)	0.033	0.057	0.178
Anonymous 6.0			
Total (TM+AI)	0.064	0.052	0.178
Technical merit (TM)	0.031	0.047	0.146
Artistic Impression (AI)	0.033	0.058	0.166
Code of points			
Total score	0.612	0.051	0.137
Grade of execution	0.164	0.060	NA
Components sum	0.295	0.062	0.114

Table 7. Results for subsamples

Each cell is a bias estimate for a particular score component and subset of the code of points sample. Coefficients are estimated using specification 3 from Table 5 and include controls for host country and home region, as well as competitor*season and round fixed effects. Standard errors (in parentheses) adjust for clustering within competitor country and competition.

	Total score	Base value	Components	GOE	Deductions	Components+GOE
By number of judges						
Fewer than 12	1.488**	0.435	0.763***	0.304*	-0.029	1.067***
	(0.646)	(0.457)	(0.277)	(0.172)	(0.049)	(0.338)
12	0.261	-0.011	0.118	0.035	-0.027	0.152
	(0.378)	(0.175)	(0.165)	(0.148)	(0.044)	(0.282)
14	0.631	0.128	0.339	0.153	-0.011	0.492
	(0.490)	(0.200)	(0.246)	(0.174)	(0.040)	(0.360)
By event						
Ice dancing	0.821***	0.023	0.685***	0.230***	-0.013	0.915***
	(0.239)	(0.077)	(0.130)	(0.071)	(0.019)	(0.186)
Ladies	0.344	-0.154	0.159*	0.170*	-0.002	0.329**
	(0.321)	(0.197)	(0.089)	(0.089)	(0.033)	(0.151)
Mens	-0.166	-0.302	0.045	-0.0001	0.072**	0.045
	(0.420)	(0.239)	(0.102)	(0.137)	(0.034)	(0.209)
Pairs	0.327	0.159	0.112	0.111	0.036	0.223
	(0.552)	(0.315)	(0.165)	(0.189)	(0.074)	(0.303)
By competition level						
Major championships	1.063***	0.0181	0.290***	0.384***	-0.024	0.674***
	(0.347)	(0.189)	(0.106)	(0.105)	(0.027)	(0.180)
Grand Prix	1.557*	0.302	0.734***	0.374*	-0.044	1.108***
	(0.814)	(0.498)	(0.230)	(0.208)	(0.075)	(0.366)
Junior events	-0.215	-0.307**	0.161**	0.0101	0.026	0.171
	(0.255)	(0.148)	(0.073)	(0.079)	(0.073)	(0.129)
By season						
2004-5	0.893*	0.102	0.362***	0.155	0.044	0.517**
	(0.504)	(0.290)	(0.128)	(0.152)	(0.051)	(0.238)
2005-6	0.736*	-0.0781	0.189	0.296*	-0.043	0.484*
	(0.436)	(0.275)	(0.140)	(0.168)	(0.046)	(0.258)
2006-7	0.296	-0.217	0.269*	0.237	-0.024	0.506*
	(0.562)	(0.289)	(0.145)	(0.185)	(0.051)	(0.275)
2007-8	0.688	0.224	0.400***	0.074	-0.014	0.474**
	(0.427)	(0.243)	(0.121)	(0.158)	(0.043)	(0.239)
2008-9	0.58	0.098	0.299***	0.084	0.001	0.383*
	(0.425)	(0.254)	(0.111)	(0.147)	(0.039)	(0.223)

Significance at the 10, 5, and 1 percent level is indicated with *, **, and ***.

Table 8. Compatriot judge effects on the max-median and inter-quartile spread of judges' scores

Each coefficient is from a separate regression (using specification 2 from Table 4 and specification 3 from Table 5) of either the max-to-median or inter-quartile spread of judges' scores on a compatriot judge indicator variable. For grades of execution (GOE) and components, the max-median and inter-quartile spreads are calculated using the sum of each judges' raw scores, unweighted by the scale of value and component factors. All regressions include controls for host country and home region, as well as fixed effects for competitor*season and rounds. Standard errors (in parentheses) adjust for clustering within competitor country and competition.

	Dependent variable	
Pre-scandal sample	Max-median	Inter-quartile
TM	-0.004	-0.005
	(0.007)	(0.010)
AI	-0.007	-0.004
	(0.008)	(0.006)
TM+AI	-0.013	-0.015
	(0.015)	(0.007)
Anonymous 6.0 sample		
TM	0.001	-0.005
	(0.009)	(0.008)
AI	-0.007	0.006
	(0.009)	(0.008)
Code of points sample		
GOE	0.116	0.006
	(0.086)	(0.030)
Components	0.012	0.077***
	(0.044)	(0.030)
GOE+Components	0.220**	0.135***
	(0.090)	(0.045)

Significance at the 10, 5, and 1 percent level is indicated with *, **, and ***.

Table 9. Interactions of judge and competitor country effects

All regressions are estimated using specification 3 from Table 5 and include controls for host country and home region, and competitor*season and round fixed effects. Standard errors (in parentheses) adjust for clustering within competitor country and competitions. In specification 1, the other disciplines variable is the number of other discplines at the same event with a compatriot judge for at least one round; this variable ranges from 0 to 3. In specification 2, the judge-country nationalistic bias is estimated using the pre-scandal sample in Zitzewitz (2006, Table 6); estimates range from 0.04 to 0.32 with mean 0.18 and SD 0.08. In specification 3, the Transparency International Index values are as reported in Zitzewitz (2006, Table 6) and range from 2 to 9.9 and has mean 5.4 and SD 2.7. In specifications 4-6, the "west bloc" in specifications 4-6 is CAN, GER, ITA, and USA and the "east bloc" is FRA, POL, RUS, and UKR. These assignments are based on estimates on the pre-scandal sample from Zitzewitz (2006, Table 8).

	Total score	Base value	Components	GOE	Components+GOE
Specification 1					
Compatriot judge in same discipline	0.316	-0.132	0.257***	0.133**	0.390***
	(0.194)	(0.109)	(0.060)	(0.059)	(0.103)
Compatriot judges in other disciplines at same event	0.128*	0.003	0.029	0.065***	0.094***
	(0.076)	(0.057)	(0.022)	(0.024)	(0.036)
Specification 2					
Compatriot judge	-0.147	-0.159	-0.0303	-0.133	-0.163
	(0.548)	(0.311)	(0.205)	(0.153)	(0.304)
(Compatriot judge)*(Judge country bias in 2001-2)	2.557	0.361	1.546***	1.596***	3.143***
	(1.528)	(0.928)	(0.553)	(0.411)	(0.881)
Specification 3					
Compatriot judge	0.578	0.248	0.468**	0.105	0.574*
	(0.564)	(0.312)	(0.188)	(0.183)	(0.321)
(Compatriot judge)*(Transparency International Index in 2001)	-0.0545	-0.0612	-0.044***	0.0037	-0.040*
	(0.045)	(0.052)	(0.015)	(0.013)	(0.023)
Specification 4					
Compatriot judge	0.123	-0.247*	0.293***	0.138*	0.431***
	(0.248)	(0.143)	(0.082)	(0.071)	(0.132)
Compatriot judge*(West Bloc competitor)	0.669	0.299	-0.0167	0.014	-0.00274
	(0.460)	(0.257)	(0.135)	(0.144)	(0.236)
Compatriot judge*(East Bloc competitor)	0.181	0.244	-0.171	-0.0646	-0.235
	(0.489)	(0.269)	(0.144)	(0.158)	(0.250)
Specification 5					
Compatriot judge	0.109	-0.254*	0.291***	0.136*	0.426***
	(0.248)	(0.143)	(0.082)	(0.071)	(0.131)
Compatriot judge*(West Bloc competitor)	0.862*	0.392	0.00168	0.106	0.108
	(0.521)	(0.281)	(0.144)	(0.159)	(0.258)
Compatriot judge*(East Bloc competitor)	0.392	0.273	-0.103	0.0473	-0.0556
	(0.507)	(0.275)	(0.155)	(0.157)	(0.258)
(West bloc judge)*(West bloc competitor)	-0.0914	0.0191	-0.0459	-0.105	-0.151
	(0.217)	(0.102)	(0.077)	(0.075)	(0.135)
(East bloc judge)*(East bloc competitor)	-0.354	-0.138	-0.0646	-0.144**	-0.209
	(0.231)	(0.124)	(0.075)	(0.073)	(0.130)
(East bloc judge)*(West bloc competitor)	-0.121	-0.0013	-0.0754	-0.0679	-0.143
	(0.197)	(0.101)	(0.063)	(0.067)	(0.113)
(West bloc judge)*(East bloc competitor)	-0.633***	-0.169	-0.187**	-0.188**	-0.375***
	(0.216)	(0.112)	(0.073)	(0.081)	(0.129)

Significance at the 10, 5, and 1 percent level is indicated with *, **, and ***.