Statistical Sleuthing by Leveraging Human Nature: A Study of Olympic Figure Skating

John W. EMERSON and Taylor B. ARNOLD

Analysis of figure skating scoring is notoriously difficult under the new *Code of Points* (*CoP*) scoring system, created following the judging scandal of the 2002 Olympic Winter Games. The CoP involves the selection of a random subpanel of judges; scores from other judges are reported but not used. An attempt to repeat the methods of previous studies establishing the presence of nationalistic bias in CoP scoring failed to recreate the competition scores from the raw scoring sheets. This raised the concern that different subpanels of judges were being selected for each skater (breaking ISU rules). However, it is also possible that the ISU was attempting to further reduce transparency in the system by permuting, separately for each skater, the order of the presentation of scores from the judging panel. Intuition suggests that it is impossible to tell the difference between accidental randomization and intentional permutation of the judges' scores. Although the recent changes do successfully prevent the study of nationalistic bias, this article provides strong evidence against the hypothesis that a separate random subpanel is chosen for each competitor. It addresses the problem by applying Gleser's extension of the Kolmogorov–Smirnov goodness-of-fit test. This article has supplementary material online.

KEY WORDS: Discrete distribution; Goodness of fit; Hypothesis test; Kolmogorov–Smirnov test.

1. INTRODUCTION

The scoring system used in international figure skating competitions changed dramatically after the 2002 Olympic judging scandal in Salt Lake City (Tibballs 2003). After awarding two gold medals in response to the scandal, the International Skating Union (ISU) introduced the new *Code of Points* (*CoP*) system and the possibility of random luck playing a role in the medal standings: the rules (International Skating Union 2008) state that the skaters' scores are calculated using a random subpanel of judges selected from the full panel at rinkside. However, all scores from the full panel of judges are published online following the competitions. Emerson (2007) showed that the silver and bronze medals would have been awarded differently at the 2006 World Championships if all judges' scores had been used, demonstrating that different subpanels favor different skaters. In similarly close competitions, the CoP will continue to award medals in a manner influenced by the random selection of subpanels.

The CoP has been widely critiqued in the popular media by skaters, experts, and independent outsiders. Some disagree as to whether certain aspects of the new system constitute strengths. For example, anonymity of judges has been touted by the ISU as helping deter corruption by reducing nationalistic pressures; others, like Zitzewitz (2010), point to this same lack of transparency as possibly increasing corruption by making it difficult for independent outsiders to look for evidence of judging bias.

Until 2010, the ISU scoring sheets preserved the anonymity of judges by presenting results in a random order common to all skaters; a given column contained the scores from the same judge for all skaters, although the identity of that judge was hidden. This allowed Emerson (2007) to study the scores of all possible subpanels and identify the particular subpanel generating the competition scores. As of 2010, this is no longer possible because a different random ordering of the judges' scores is presented for each skater. For any given skater it is possible to study all possible subpanels and identify the subpanel (or possibly a few subpanels) generating the competition scores, but scores of the subpanels cannot be compared across skaters. Thus, we cannot say, "a particular subpanel of judges favored one skater over another."

What is behind this change in 2010? A troubling explanation is that the computerized scoring system is selecting a different subpanel for each skater instead of using the same subpanel across all skaters. We refer to this as accidental randomization of the judges' scores, which would violate the ISU's own competition rules (International Skating Union 2008).

However, an alternative explanation is that the ISU chose consistent judging subpanels but decided to intentionally permute the presented judges' scores to hinder further studies of nationalistic bias. The ISU insists that judging anonymity is a

John W. Emerson is Associate Professor (E-mail: *john.emerson@yale.edu*) and Taylor B. Arnold is Doctoral Candidate (E-mail: *taylor.arnold@yale.edu*), Department of Statistics, Yale University, P.O. Box 208290, New Haven, CT 06520-8290. The authors thank Ankur Sharma for his help validating the raw data scrape, and are grateful for several helpful discussions arising from seminars presenting parts of this work. The authors also thank the Associate Editor and two referees for their helpful comments.

 DOI: 10.1198/tast.2011.10165

critical part of the system; without this, judges could be pressured to reward particular skaters and/or penalize their competition. When the CoP was introduced, the ISU argued that the anonymous reporting of scores and the random selection of subpanels of judges might reduce the likelihood of collusion among judges by making it impossible to verify whether the promised (or implied) vote-trading actually occurred. In 2010, perhaps the ISU felt that the intentional permutation of scores would further improve the system or eliminate negative publicity by impeding detailed analyses. This change would manifest itself in exactly the same way as accidental randomization of the scores, but prevents the analysis of subpanel results across skaters. Intuition suggests that it is impossible to tell the difference between accidental randomization and intentional permutation of the judges' scores.

This article starts by considering the implications of a "broken" system where different subpanels of judges are used to score different competitors, which represents the accidental randomization (null) hypothesis. Several metrics can help detect differences in judging patterns between subpanels and excluded judges. For example, one metric is the difference between the mean executed element scores of the subpanel and the excluded judges. Under the subpanel selection mechanism specified by the null hypothesis and conditional on the actual scores of the competition, the value of the metric for each competitor is randomly selected from the collection of values of the metric formed by considering all possible subpanels. If, on the other hand, the system appropriately uses the same subpanel across all skaters in a given event, differences in average scoring tendencies between the subpanel and the excluded judges could be apparent in the calculated metrics. Other aspects of human nature, such as variability in scoring, are captured by other metrics. The Kolmogorov–Smirnov goodness-of-fit test, modified by Conover (1972) (one-sided) and Gleser (1985) (two-sided) for discontinuous distributions and adapted for the purpose of this particular problem, is the basis for the analysis.

Section 2 presents the data, Section 3 describes the methodology, and Section 4 presents the results of the analysis of the 2010 Olympic Winter Games and studies the power of the methodology in the context of the 2009 European Championships. It also discusses the improvements to Kolmogorov–Smirnov goodness-of-fit testing in the R Language and Environment for Statistical Computing (R Development Core Team 2011). A byproduct of this work, R package dgof (Arnold and Emerson 2011), is provided to support the methodology. Data and other materials relating to the figure skating competitions are available as supplementary material.

2. THE DATA

This article considers data from the 2009 European Championships, where judges' scores were reported reliably column-by-column across all skaters, and the 2010 Olympic Winter Games, where judges scores were permuted skater-by-skater either accidentally or intentionally. International skating competitions include singles events for men and ladies as well as pairs and ice dancing. (Because ice dancing involves three program segments and is quite different from the other events, it will not be studied here.) The singles and pairs events consist of short program and free skate segments, with some competitors eliminated from the competition as a result of the short program segment. Between 20 and 30 competitors competed in each segment of the 2010 Olympic Winter Games.

Figure 1 shows the official scoring sheet from the Men's Free Skate segment of 2010 Olympic silver medallist Evgeni

Rank	Name	NOC Code	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
2	PLUSHENKO Evgeni	RUS	24	165.51	82.71	82.80	0.00

#	Executed Elements	Info	Base Value	GOE	The Judges Panel (in random order)									Scores of Panel
1	4T+3T		13.80	0.80	1	1	1	1	0	-1	1	1	2	14.60
2	3A		8.20	-0.36	0	-1	1	0	-1	-1	-1	0	1	7.84
3	3A+2T		9.50	1.00	1	2	1	1	0	0	2	1	2	10.50
4	3Lo		5.00	0.60	0	0	1	1	-1	0	1	1	1	5.60
5	FSSp3		2.60	0.14	0	0	2	1	-1	-1	2	0	1	2.74
6	3Lz		6.00	0.60	0	1	1	1	0	0	2	0	2	6.60
7	CSSp4		3.00	0.70	1	2	2	1	1	1	2	1	2	3.70
8	CiSt3		3.30	0.80	2	1	3	2	0	2	2	1	2	4.10
9	3Lz+2T		8.03	x 0.00	0	0	1	0	-1	-1	-1	0	1	8.03
10	3S		4.95	x 0.80	1	1	1	1	0	1	1	0	1	5.75
11	2A		3.85	x 1.00	0	1	1	1	0	2	2	1	1	4.85
12	SlSt3		3.30	1.00	2	1	3	3	1	3	2	1	2	4.30
13	CCoSp4		3.50	0.60	1	1	2	1	1	1	2	2	1	4.10
			75.03											**82.71**
	Program Components			**Factor**										
	Skating Skills			2.00	9.00	7.75	9.00	9.00	8.00	7.25	8.00	8.25	9.00	8.40
	Transitions / Linking Footwork			2.00	8.75	6.00	8.00	8.75	6.00	6.50	7.25	7.25	8.50	7.25
	Performance / Execution			2.00	9.00	9.50	9.50	8.75	8.25	8.00	8.75	8.25	9.25	8.80
	Choreography / Composition			2.00	9.25	7.75	9.00	9.00	7.75	7.50	8.50	7.75	8.75	8.20
	Interpretation			2.00	9.25	9.50	9.50	9.00	7.50	7.75	8.50	8.00	9.50	8.75
	Judges Total Program Component Score (factored)													**82.80**
	Deductions:													**0.00**

Figure 1. Evgeni Plushenko's 2010 Olympic free skate results.

Table 1. Scoring example: the top row reproduces the entries of Plushenko's second executed element, a triple Axel (3A), shown in Figure 1. The judges' marks (−1, 0 or 1) are mapped to a scale of values using the SOV table presented as supplementary material. The segment subpanel excluded the first and seventh judges; a trimmed mean of the remaining five values (bottom) produced a deduction of 0.36 from the base value 8.20, resulting in the panel score 7.84.

Base value	GOE	The judges panel (in random order)									Scores of panel
8.20	−0.36	0	−1	1	0	−1	−1	−1	0	1	7.84
Scale of values		0.0	−1.4	1.0	0.0	−1.4	−1.4	−1.4	0.0	1.0	
Excluded		X						X			
Trimmed			X	X							
Averaged					0.0	−1.4	−1.4		0.0	1.0	⇒ −0.36
											+8.20
											= 7.84

Plushenko. Plushenko and gold medalist Evan Lysacek received exactly the same *program component* scores—generally reflecting artistry—while Lysacek received higher scores than Plushenko for the quality of his *executed elements*. The program component scores can range from 0.00 to 10.00 in increments of 0.25, and the *scores of panel* in the right column are the result of a trimmed mean of the seven judges' scores on the scoring subpanel. In the men's free skate, the trimmed means are then multiplied by a factor of two to achieve the desired weight in the total score, which explains the magnitude of 82.80 for Plushenko's program component total.

The executed elements involve a more complicated calculation, starting with a *base value* reflecting the difficulty of the element. Note the extremely high base value, 13.80, of Plushenko's unmatched quadruple-toe/triple-toe combination, his first executed element. An "x" appearing to the right of a base value indicates a 10% bonus for jumps performed in the second half of the program, as with Plushekno's executed elements 9–11. For each of the executed elements, the judges provide integer marks from −3 to 3, with zero reflecting an average quality of execution. These marks are then transformed in a way that depends on the degree of difficulty, and combined to produce the total *grade of execution* (*GOE*) in the same way described for the program component scores, using a trimmed mean of the seven judges' scores on the scoring subpanel. The mappings needed for these values appear in a *scale of values* (*SOV*) table, included as supplementary material.

For example, consider Plushenko's second executed element, a triple Axel. The first row of Table 1 shows the base value, average grade of execution, judge evaluations, and the total score, which also are in the second row of Figure 1. The total score 7.84 is the base value 8.40 plus the average grade of execution, a penalty of −0.36 in this case for slightly below-average perceived quality. The entries in the next row of Table 1 reflect the appropriate scale of values for a triple Axel; the scores in columns 1 and 7 were excluded by random selection. After these exclusions, one each of the −1.4 and 1.0 values are trimmed. The average grade of execution −0.36 is the mean of the remaining five values shown in the bottom row of Table 1.

Plushenko's free skate results provide an example of the fundamental unit of information central to this article: the scores of all judges for the performance of a skater in a given segment of the competition. For each executed element or program component (e.g., each row of a scoring sheet), there are likely several subpanels of seven judges whose scores reproduce the observed competition score. For example, it can be seen in Table 1 that a subpanel excluding judges in columns 4 and 5 would also produce the observed panel score, 7.84. However, only by excluding the judges in columns 1 and 7 can the complete set of panel scores, shown in the rightmost column of Figure 1, be obtained. In some rare cases (i.e., four times in 148 performances of the 2010 Olympic Winter Games), more than one subpanel could have generated the same competition scores; we omit these four performances from the analysis.

3. METHODOLOGY

Let the null hypothesis correspond to the accidental randomization theory, with different judging subpanels selected at random for each skater. The alternative hypothesis represents intentional permutation of the presented scores, with a common randomly selected judging subpanel used throughout each competition segment as described in the ISU rules. Our hypothesis test is conditional on the observed competition scores. The rejection of the null hypothesis depends on patterns of human nature evident in judging panels which would not be detectable under the null hypothesis. Suppose, for example, that the system is working as intended (the alternative hypothesis) and that the two excluded judges tend to give lower scores than the judges on the selected subpanel. In this case, subtracting the mean scores of the two excluded judges from the mean scores of the subpanel would tend to produce larger results across all skaters in the event than would be expected under accidental randomization (the null hypothesis).

Consider a particular competition segment having S competitors. For each skater s (or duo in a pairs event) there are $\binom{9}{7} = 36$ possible partitions of the matrix of scores into two submatrices corresponding to a subpanel of seven judges and the excluded two judges, denoted x_s^i and y_s^i, respectively, for the ith partition. There is at least one, and almost always exactly one, such partition corresponding to the actual scoring subpanel of the competition; let x_s^* denote the submatrix of scores of this actual scoring subpanel, and let y_s^* denote the matrix of scores of the two excluded judges.

Let $m(x_s^i, y_s^i)$ be a metric capturing some element of contrast between a subpanel and its associated pair of excluded judges. For example, m might equal the difference between the means of all executed element scores (or, alternatively, the program

component scores) from the two groups; we discuss other metrics used in the analysis later in this section. For each of the 36 possible partitions, calculate $m(x_s^i, y_s^i)$; cases where two or more of these are equal could either be omitted from the study or a small jittering could break the ties without disrupting the order with neighboring values (producing 36 unique values). Let $\widehat{F}_s$ represent the empirical cumulative distribution function obtained from these 36 values.

Let $\widehat{F}_s^*$ be the proportion of values of the metric less than or equal to the value associated with the actual scoring panel. Alternatively, $\widehat{F}_s^*$ may be described as the empirical cumulative probability of the observed value of the metric, which lies in the set $\Omega = \{1/36, 2/36, \ldots, 36/36\}$:

$$\widehat{F}_s^* \equiv \widehat{F}_s(m(x_s^*, y_s^*)) = \frac{\#\{m(x_s^i, y_s^i) \le m(x_s^*, y_s^*)\}}{36}, \quad (1)$$

where $i \in \{1, 2, \ldots, 36\}$. Under the null hypothesis, the distribution of $\widehat{F}_s^*$ is a discrete uniform distribution,

$$\widehat{F}_s^* \overset{\text{iid}}{\sim} \text{Uniform}(\Omega) \quad (2)$$

for $s \in \{1, 2, \ldots, S\}$. Using the S observed cumulative empirical probabilities, our methodology applies the Kolmogorov type goodness-of-fit test presented in Gleser (1985) to obtain a p-value $p_{e,m}^*$ for the event segment e using the chosen metric m. The process may be repeated separately for each competition segment and for each of several metrics reflecting different aspects of human nature in judging.

Under the alternative hypothesis of intentional permutation, one particular subpanel is used to score all competitors in a particular competition segment. In this case, the values of the metrics corresponding to the actual competition subpanel ($m(x_s^*, y_s^*)$) may exhibit nonuniform patterns. This subpanel might, for example, exclude two particularly enthusiastic judges who generally award scores higher than the subpanel of seven judges. In such a case, the values of $m(x_s^*, y_s^*)$ would be among the lowest of all possible values of $m(x_s^i, y_s^i)$; the resulting S values of $\widehat{F}_s^*$ would generally be small. As a result, the one-sample Gleser–Kolmogorov–Smirnov test should detect a departure from independence and uniformity of the distribution of the $\widehat{F}_s^*$, producing a smaller p-value $p_{e,m}^*$.

Finally, conditional on the observed judges' scores, the results of all six competition segments provide independent information which greatly increases the power of the testing procedure. For a specified significance level α [e.g., perhaps 0.05/6 per Bonferroni, or a similar value following the method proposed by Šidàk (1967) or the method proposed by Westfall and Wolfinger (1997)], we reject the null hypothesis if *any* of the six tests exceed this threshold. This provides a Type I error rate of about $1 - (1 - \alpha)^6$, or 0.049 if $\alpha = 0.05/6$.

Competitions prior to 2010 provide the opportunity to study scores judge-by-judge across skaters, allowing an exploration of the power of the methodology. For example, in the Ladies Short Program of the 2009 European Championships the judges with scores presented in columns 2 and 7 were omitted, and the other seven judges formed the scoring subpanel. Of course, there were 35 other subpanels which could have been selected in the competition; these 36 possible subpanels constitute the full extent of the alternative hypothesis under ISU rules.

For each such subpanel in each of the six event segments, we apply the methodology described above and tabulate the rejections of the null hypothesis. That is, for event segment e,

$$1 - \beta_{e,m} = \frac{\#\{p_{e,m}^i \le 0.05/6\}}{36},$$

where $i \in \{1, 2, \ldots, 36\}$ indexes the possible panels of the alternative hypothesis. For a given metric, we reject the null hypothesis when *any* of the six events lead to rejecting the null hypothesis. Thus, the power achieved by aggregating over the six event segments is

$$1 - \prod_e \beta_{e,m}$$

for metric m.

We consider six metrics capturing aspects of human nature that might manifest themselves when comparing a judging subpanel to the corresponding two excluded judges. The first two metrics, *MeanEE* and *MeanPC*, were alluded to previously; they simply calculate the difference between the mean scores of the two sets of judges for the executed elements and the program component scores, respectively. The next two, *LowExtr* and *HighExtr*, focus on extreme scores respectively, counting the number of times scores of the excluded two judges exceed minimum or maximum scores of the subpanel, separately for each executed element or program component. For example, consider the scoring sheet for Plushenko shown in Figure 1, and suppose the judging subpanel consists of the first seven columns of the judging panel. On the first executed element (4T + 3T), one of the excluded judges provided a rating (2) that exceeded the maximum rating of the subpanel (1), but this is the only such occurrence in Plushenko's free skate scores. Thus, the value of *HighExtr* for Plushenko and this particular subpanel would be 1. The final two metrics, *VarDiffEE* and *VarDiffPC*, use the difference between the average variances of executed element scores of the two groups, and the difference between the average variances of program component scores. With the program component scores in Figure 1 and again considering the judging subpanel corresponding to the first seven columns, *VarDiffPC* would be calculated by first obtaining the variances of the program component scores of each of the nine judges (0.044, 2.144, 0.375, 0.019, 0.781, 0.331, 0.356, 0.175, 0.156), then averaging them for the subpanel (0.579) and for the excluded judges (0.166), and finally taking the difference (0.413). Thus, a large value of *VarDiffPC*, for example, corresponds to a case where the excluded judges illustrate far less variability in the program component scores than the judging subpanel.

4. RESULTS AND DISCUSSION

The application of the methodology presented in Section 3 provides compelling evidence that the scoring system used in the 2010 Olympic Winter Games simply permuted results in the scoring sheets, hiding possible evidence of nationalistic bias. Table 2 presents the full results; a small amount of jittering was used to break ties, discussed earlier, and had no substantive impact on the results. Striking patterns are evident in most of the events with most of the metrics. The only metric failing

Table 2. Analysis of 2010 Olympic Winter Games: p-values from goodness-of-fit tests indicating that the distribution of values of various metrics is not uniform (as would be the case under the null hypothesis).

		Metric					
Event	Segment	*LowExtr*	*HighExtr*	*VarDiffEE*	*VarDiffPC*	*MeanEE*	*MeanPC*
Ladies	Short Pr.	0.96383	0.82852	0.15242	0.11292	0.33512	0.76238
	Free Sk.	0.00012	0.01469	0.10203	0.00001	0.02254	0.00022
Men	Short Pr.	0.01113	0.10797	0.03451	0.00001	0.07941	0.00003
	Free Sk.	0.00040	0.00006	0.19241	0.00040	0.01469	<0.00001
Pairs	Short Pr.	0.00011	<0.00001	0.01311	0.10156	0.00093	0.00093
	Free Sk.	0.00327	0.19882	0.19882	0.09949	0.00475	0.50883

to reject the null hypothesis with our methodology was *VarDiffEE*; however, the results presented for *VarDiffEE* in Table 2 could be seen as weak evidence against the null hypothesis. In the aggregate, the evidence is overwhelming, indicating that the selection of judges was most certainly not repeated separately for each competitor. Very similar results were obtained using Cramér von-Mises tests instead of the Kolmogorov type tests of Conover (1972) and Gleser (1985).

The Ladies Short Program is the one event segment without sufficient evidence to reject the null hypothesis on its own; evidence from the other event segments clearly point to the system working as designed by the ISU. In this case then, the panel selection may have divided the judges into groups with similar characteristics, or differing in characteristics not captured by our metrics. This leads to an unusual observation about the proposed methodology: if, in the extreme case, judges produced scores in the manner of independent, identically distributed robots, the lack of patterns relating to human nature in scoring would make it impossible for this methodology to distinguish between the two hypotheses.

Results from the exploration of power using the 2009 European Championships are shown in Table 3. The primary six rows summarize the proportions $1 - \beta_{e,m}$ of 36 subpanels for each event and metric which would have led to rejecting the null hypothesis at significance level $\alpha = 0.05/6$. *VarDiffPC* is the most powerful metric. For any one event segment and choice of metric, the powers are unimpressive because of the relatively small sample sizes (numbers of competitors in an event segment). However, combining the results of the six events for each metric provides strong overall power. Notice that while the analysis of events can be combined for a given metric (the selection of panels is independent across events), the results within an event across different metrics are correlated.

The power results from the 2009 European Championships are indicative of the value of the methodology, although there is no guarantee of similar power in other competitions. While we would expect to see patterns in the scoring of international skating events, differences in the judges selected to participate and the overall atmosphere of the Olympic Games, for example, could influence the judges in different ways. One difference between the competitions studied here is with the *MeanPC* metric. The 2009 dataset suggests it has relatively poor power, but in the 2010 data it gives one of the lowest sets of p-values. This might simply be good luck in this case, when the evidence against the null appears overwhelming, or it might be related to unobservable differences the nature of judging these competitions.

Previous studies have focused on nationalistic bias in judging—one undesirable aspect of human nature, whether intentional or subconscious. However, nationalistic bias is not a useful aspect of human nature for answering the question addressed in this study because it does not, by definition, create patterns evident across most or all of the skaters. This study relies on other aspects of human nature to address the question of interest: is the selection of scoring subpanels working as intended? The evidence shows that in 2010 the ISU intentionally started permuting the presented judges' scores, skater by skater, making it far more difficult—perhaps even impossible—to repeat the analyses of Zitzewitz (2010), Emerson (2007), or to attempt to modify the Olympic diving analysis of Emerson, Seltzer and Lin (2009) for figure skating. The change does not violate ISU rules but does reduce transparency in the scoring system, helping to obscure aspects of the system which may

Table 3. Power exploration based on analysis of 2009 European Championship results.

		Metric					
Event	Segment	*LowExtr*	*HighExtr*	*VarDiffEE*	*VarDiffPC*	*MeanEE*	*MeanPC*
Ladies	Short Pr.	0.278	0.333	0.361	0.389	0.167	0.361
	Free Sk.	0.194	0.028	0.222	0.750	0.472	0.000
Men	Short Pr.	0.056	0.306	0.222	0.333	0.194	0.222
	Free Sk.	0.472	0.444	0.083	0.694	0.417	0.167
Pairs	Short Pr.	0.278	0.250	0.333	0.500	0.361	0.250
	Free Sk.	0.028	0.167	0.306	0.472	0.139	0.000
Overall power		0.796	0.844	0.836	0.992	0.886	0.689

not be in the best interest of skaters and which threaten to embarrass the ISU.

The R Language and Environment for Statistical Computing (R Development Core Team 2011) was used for this study and provides Kolmogorov–Smirnov goodness-of-fit tests in the core **stats** package. It permits specification of a discontinuous null distribution for one-sample tests, but the algorithm is not well suited for calculating the test statistic in such cases and particularly in small-sample situations with ties in the data (typically expected with discontinuous distributions). This article offers package dgof, described in Arnold and Emerson (2011), containing a proposed revision of R's `ks.test()` function, which offers an improved implementation addressing this shortcoming and adding the method proposed by Conover (1972) and refined by Gleser (1985) for one-sample goodness-of-fit tests with discrete distributions. The calculations of exact p-values are only provided for samples size of at most 30 because of numerical precision challenges. With sample sizes >30 and <100 and no ties in the data, exact p-values are used following the methodology of Marsaglia, Tsang and Wang (2003) for the two-sided case or Birnbaum and Tingey (1951) for the one-sided case. With sample sizes ≥ 100, or in cases with intermediate sample sizes having ties in the data, the standard asymptotic Kolmogorov–Smirnov p-values are used, which are known to be conservative for noncontinuous null distributions (Slakter 1965).

SUPPLEMENTARY MATERIALS

Data: Figure skating results from the 2009 European Championships and the 2010 Olympic Winter Games. Both raw scoring sheets (PDF) and processed files (CSV) are available, along with the most recent scale of values table (CSV). (EmersonArnold_Data.zip)

R-package dgof: R package **dgof** (Arnold and Emerson 2011) containing function `ks.test()`, a proposed modification to the function by the same name in R's recommended package **stats**. This package is currently available from CRAN: *http://cran.r-project.org/web/packages/dgof/*.

[Received August 2010. Revised July 2011.]

REFERENCES

Arnold, T. B., and Emerson, J. W. (2011), "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions," *The R Journal*, to appear. [144,148]

Birnbaum, Z. W., and Tingey, F. H. (1951), "One-Sided Confidence Contours for Probability Distribution Functions," *The Annals of Mathematical Statistics*, 22 (4), 592–596. [148]

Conover, W. J. (1972), "A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions," *Journal of the American Statistical Association*, 67 (339), 591–596. [144,147,148]

Emerson, J. W. (2007), "Chance, on and off the Ice," *Chance*, 20 (2), 19–21. [143,147]

Emerson, J. W., Seltzer, M., and Lin, D. (2009), "Assessing Judging Bias: An Example From the 2000 Olympic Games," *The American Statistician*, 63 (2), 124–131. [147]

Gleser, L. J. (1985), "Exact Power for Goodness-of-Fit Tests of Kolmogorov Type for Discontinuous Distributions," *Journal of the American Statistical Association*, 80 (392), 954–958. [144,146-148]

International Skating Union (2008), "Special Regulations & Technical Rules: Single & Pair Skating and Ice Dance 2008," as accepted by the 52nd Ordinary Congress of the International Skating Union, June 2008. [143]

Marsaglia, G., Tsang, W. W., and Wang, J. (2003), "Evaluating Kolmogorov's Distribution," *Journal of Statistical Software*, 8 (18), available at *http://www.jstatsoft.org/v08/i18/*. [148]

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available at *http://www.R-project.org/* [144,148]

Šidàk, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," *Journal of the American Statistical Association*, 62 (318), 626–633. [146]

Slakter, M. J. (1965), "A Comparison of the Pearson Chi-Square and Kolmogorov Goodness-of-Fit Tests With Respect to Validity," *Journal of the American Statistical Association*, 60 (311), 854–858. [148]

Tibballs, G. (2003), *Great Sporting Scandals: From Over 200 Years of Sporting Endeavours*, London: Robson Books. [143]

Westfall, P. H., and Wolfinger, R. D. (1997), "Multiple Tests With Discrete Distributions," *The American Statistician*, 51, 3–8. [146]

Zitzewitz, E. (2010), "Does Transparency Really Increase Corruption? Evidence From the Reform of Figure Skating Judging," unpublished manuscript, Dartmouth College. [143,147]