The Effect of Transparency on Subjective Evaluations

Evidence from Competitive Figure Skating

Chui Yee Ho and Ximeng Fang

EEA-ESEM Rotterdam, 29 August 2024

Evaluations and decisions are often made by groups

High-stakes decisions under uncertainty are often delegated to **groups of evaluators** rather than single individuals

- e.g. juries, expert panels, hiring committees, peer review, ...
- the study of collective intelligence has a long-standing scientific tradition (e.g. Condorcet 1785, Galton 1907)

Evaluations and decisions are often made by groups

High-stakes decisions under uncertainty are often delegated to **groups of evaluators** rather than single individuals

- e.g. juries, expert panels, hiring committees, peer review, ...
- the study of collective intelligence has a long-standing scientific tradition (e.g. Condorcet 1785, Galton 1907)

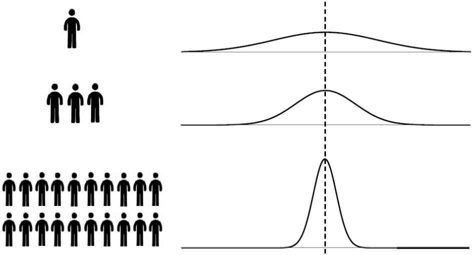


Figure 1: Normal probability distributions of errors for an individual judgment and collective (average) judgments by three and twenty individuals

But crowds are not necessarily wise

$$\textit{Evaluation} = \textit{``true'' value} + \textit{bias} + \textit{noise}$$

But crowds are not necessarily wise

$$\textit{Evaluation} = \textit{``true'' value} + \textit{bias} + \textit{noise}$$

The **accuracy of group decisions** is constrained by how individuals members form and report their judgments

- risk of low effort when trying to find the "true value"
- systematic biases may not average out even in large groups
- herding and groupthink can create correlated errors

But crowds are not necessarily wise

$$\textit{Evaluation} = \textit{“true” value} + \textit{bias} + \textit{noise}$$

The **accuracy of group decisions** is constrained by how individuals members form and report their judgments

- risk of low effort when trying to find the “true value”
- systematic biases may not average out even in large groups
- herding and groupthink can create correlated errors

How do institutional features affect evaluation decisions in groups?

- One important feature: Are opinions of individual members made **transparent**? (Prat, 2005; Levy, 2007; Gersbach/Hahn, 2012; Fehrler/Hughes, 2018; Mattozzi/Nakaguma, 2019; Fehrler/Janas, 2021; Benesch et al., 2018; Hansen et al., 2018)

Our study: Evaluation of competitive sports performances

We study the **effect of a transparency reform** to the judging system for figure skating competitions

Figure skating scores are awarded by a panel of judges

A panel of (nine) judges evaluates both the technical execution and the artistic value of a skater's performance

Figure skating scores are awarded by a panel of judges

A panel of (nine) judges evaluates both the technical aspects and the artistic value of a skater's performance

- Technical elements score: difficulty and execution of technical elements (e.g. jumps, spins)
- Program component score: more artistic aspects of the performance (e.g. choreography, expressiveness, ...)

Figure skating scores are awarded by a panel of judges

A panel of (nine) judges evaluates both the technical aspects and the artistic value of a skater's performance

- Technical elements score: difficulty and execution of technical elements (e.g. jumps, spins)
- Program component score: more artistic aspects of the performance (e.g. choreography, expressiveness, ...)

The total score is computed by **averaging the individual judges' scores** (trimmed by the highest and lowest scores)

- Judge submits their score independently from each other
- Communication is not allowed

Figure skating has seen its share of judging scandals

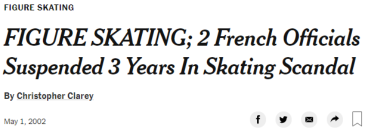


FIGURE SKATING

FIGURE SKATING; 2 French Officials Suspended 3 Years In Skating Scandal

By Christopher Clarey

May 1, 2002

Figure skating has seen its share of judging scandals

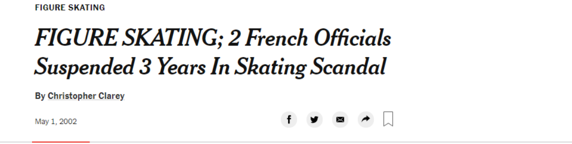
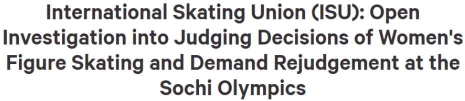


FIGURE SKATING

FIGURE SKATING; 2 French Officials Suspended 3 Years In Skating Scandal

By Christopher Clarey

May 1, 2002



International Skating Union (ISU): Open Investigation into Judging Decisions of Women's Figure Skating and Demand Rejudgement at the Sochi Olympics

Figure skating has seen its share of judging scandals

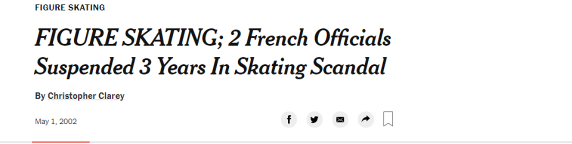
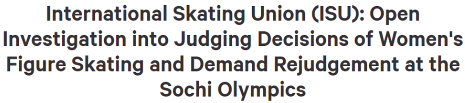


FIGURE SKATING

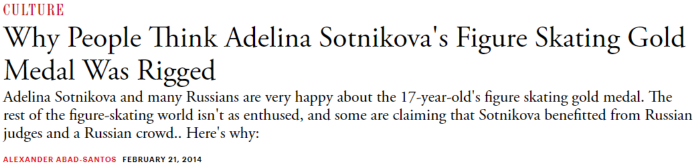
FIGURE SKATING; 2 French Officials Suspended 3 Years In Skating Scandal

By Christopher Clarey

May 1, 2002



International Skating Union (ISU): Open Investigation into Judging Decisions of Women's Figure Skating and Demand Rejudgement at the Sochi Olympics



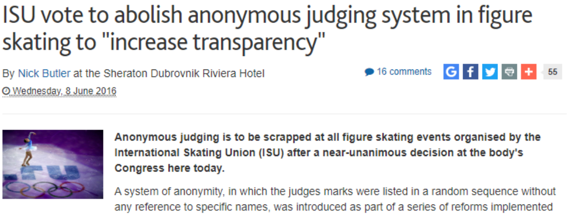
CULTURE

Why People Think Adelina Sotnikova's Figure Skating Gold Medal Was Rigged

Adelina Sotnikova and many Russians are very happy about the 17-year-old's figure skating gold medal. The rest of the figure-skating world isn't as enthused, and some are claiming that Sotnikova benefitted from Russian judges and a Russian crowd.. Here's why:

ALEXANDER ABAD-SANTOS FEBRUARY 21, 2014

Transparency reform in the publication of scores in 2016



ISU vote to abolish anonymous judging system in figure skating to "increase transparency"

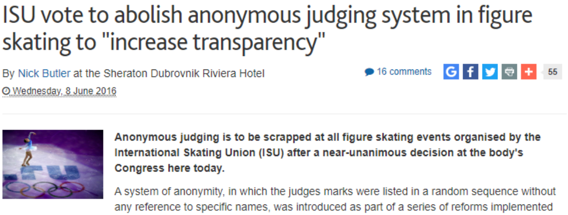
By Nick Butler at the Sheraton Dubrovnik Riviera Hotel

Wednesday, 8 June 2016

Anonymous judging is to be scrapped at all figure skating events organised by the International Skating Union (ISU) after a near-unanimous decision at the body's Congress here today.

A system of anonymity, in which the judges marks were listed in a random sequence without any reference to specific names, was introduced as part of a series of reforms implemented

Transparency reform in the publication of scores in 2016



ISU vote to abolish anonymous judging system in figure skating to "increase transparency"

By Nick Butler at the Sheraton Dubrovnik Riviera Hotel

Wednesday, 8 June 2016

16 comments

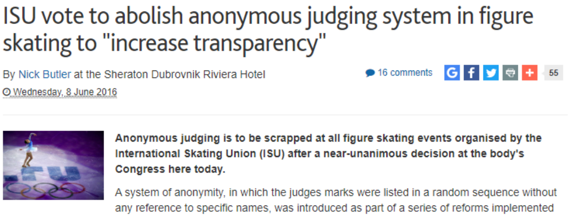
55

Anonymous judging is to be scrapped at all figure skating events organised by the International Skating Union (ISU) after a near-unanimous decision at the body's Congress here today.

A system of anonymity, in which the judges marks were listed in a random sequence without any reference to specific names, was introduced as part of a series of reforms implemented

- Pre-reform: anonymized publishing of individual scores without link to judge identity

Transparency reform in the publication of scores in 2016



ISU vote to abolish anonymous judging system in figure skating to "increase transparency"

By Nick Butler at the Sheraton Dubrovnik Riviera Hotel

Wednesday, 8 June 2016

Anonymous judging is to be scrapped at all figure skating events organised by the International Skating Union (ISU) after a near-unanimous decision at the body's Congress here today.

A system of anonymity, in which the judges marks were listed in a random sequence without any reference to specific names, was introduced as part of a series of reforms implemented

- Pre-reform: anonymized publishing of individual scores without link to judge identity
- Post-reform (2016/17 season onwards): scores by each judge in the panel are made public

“Beauty contest” model as theoretical framework

Model of (strategic) evaluation building on Morris/Shin (2002):
Judge j observes a performance, evaluates its quality, reports score π_j

- judge exerts effort $\tau_j > 0$ to generate a signal $x_j = \theta + \epsilon_j$
- “true” quality θ (with common prior: $\mathcal{N}(\mu, \sigma^2)$)
- noise term $\epsilon_j \sim \mathcal{N}(\mu, \frac{\sigma^2}{\tau_j})$

“Beauty contest” model as theoretical framework

Model of (strategic) evaluation building on Morris/Shin (2002): Judge j observes a performance, evaluates its quality, reports score π_j

- judge exerts effort $\tau_j > 0$ to generate a signal $x_j = \theta + \epsilon_j$
- “true” quality θ (with common prior: $\mathcal{N}(\mu, \sigma^2)$)
- noise term $\epsilon_j \sim \mathcal{N}(\mu, \frac{\sigma^2}{\tau_j})$

After Bayesian updating, the **judge reports the score** π_j that maximizes the expectation of

$$U_j(\pi, \tau_j, \theta) = -(\pi_j - \theta \underbrace{-b_j}_{\text{bias}})^2 - \underbrace{\eta\ (\pi_j - \bar{\pi}_{\text{-}j})^2}_{\text{“conformity” motive}} - \underbrace{c\,\tau_j}_{\text{effort cost}}$$

Theoretical predictions for the effects of transparency

The model generates **several predictions** that are also empirically testable using our data:

Theoretical predictions for the effects of transparency

The model generates **several predictions** that are also empirically testable using our data:

1. $\frac{\partial}{\partial \eta} Var[\pi_j|\theta] < 0$: score dispersion within the judge panel decreases.
    - ▶ scores become more similar
    - three channels: higher effort, more conservatism, bias-matching

Theoretical predictions for the effects of transparency

The model generates **several predictions** that are also empirically testable using our data:

1. $\frac{\partial}{\partial\eta} Var[\pi_j|\theta] < 0$: score dispersion within the judge panel decreases.
   - ▶ scores become more similar
   - three channels: higher effort, more conservatism, bias-matching
2. $\frac{\partial^2}{\partial\eta\,\partial\sigma} Var[\pi_j|\theta] < 0$: effect increases with subjectivity.
   - e.g., artistic versus technical score

Theoretical predictions for the effects of transparency

The model generates **several predictions** that are also empirically testable using our data:

1. $\frac{\partial}{\partial \eta} Var[\pi_j|\theta] < 0$: score dispersion within the judge panel decreases.
    - ▶ scores become more similar
    - three channels: higher effort, more conservatism, bias-matching
2. $\frac{\partial^2}{\partial \eta\, \partial \sigma} Var[\pi_j|\theta] < 0$: effect increases with subjectivity.
    - e.g., artistic versus technical score
3. $\frac{\partial^2}{\partial \eta\, \partial b_j} E[\pi_j|\theta] = 0$: no decrease in the *aggregate* bias.
    - judges try to match each others' biases
    - ▶ individual effects cancel each other out

Empirical strategy to identify effects of transparency

Study the effect of transparency using the 2016 figure skating reform.

Empirical strategy to identify effects of transparency

Study the effect of transparency using the 2016 figure skating reform.

- Junior Grand Prix (JGP) events already published judge scores openly prior to the reform → use as control group

Empirical strategy to identify effects of transparency

Study the effect of transparency using the 2016 figure skating reform.

- Junior Grand Prix (JGP) events already published judge scores openly prior to the reform → use as control group
- "Treated" events: all Senior events (e.g. Olympics, Grand Prix, Championships) and other Junior events → Non-JGP events

Empirical strategy to identify effects of transparency

Study the effect of transparency using the 2016 figure skating reform.

- Junior Grand Prix (JGP) events already published judge scores openly prior to the reform → use as control group
- "Treated" events: all Senior events (e.g. Olympics, Grand Prix, Championships) and other Junior events → Non-JGP events

Difference-in-differences design: compare changes in judge scores

Empirical strategy to identify effects of transparency

Study the effect of transparency using the 2016 figure skating reform.

- Junior Grand Prix (JGP) events already published judge scores openly prior to the reform → use as control group
- "Treated" events: all Senior events (e.g. Olympics, Grand Prix, Championships) and other Junior events → Non-JGP events

Difference-in-differences design: compare changes in judge scores

- ideally want to know each judge's scores, but anonymous judging pre-reform!

▶ analyze distribution of scores in the judge panel

Data on performances and scores

Data on figure skating competitions from seasons 2013-14 to 2019-20 obtained by scraping the official ISU website (www.isu.org):

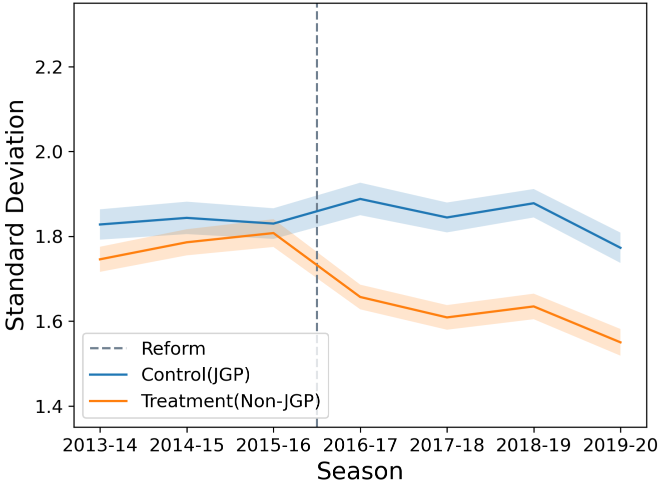
- info on scores as well as skater and judge identities
- can identify "compatriot" performances

		JGP (control)		Non-JGP (treated)	
	full sample	pre-reform	post-reform	pre-reform	post-reform
# Performances	16821	3103	4340	3994	5384
# Rounds	1028	152	200	292	384
# Events	127	21	28	34	44
# Skaters/athletes	1905	711	954	617	730
# Judges	563	333	379	323	338

Effects on score dispersion

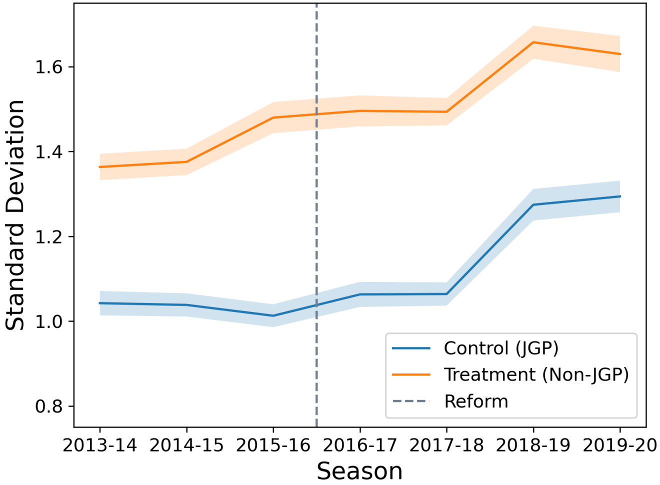
Decrease in the standard deviation of artistic scores

Figure 1: SD of the artistic score within the judge panel



But not for the more objective technical scores

Figure 2: SD of the technical score within the judge panel



Effect on within-panel standard deviation

Table 1: Estimated effect of transparency on score dispersion

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	-0.014	-0.033	0.008	-0.018	-0.009
	(0.041)	(0.043)	(0.020)	(0.021)	(0.020)
Post × Non-JGP	-0.121***	-0.103**	-0.025	-0.034	-0.009
	(0.045)	(0.049)	(0.028)	(0.028)	(0.029)
Skater FEs	—	Yes	—	Yes	Yes
Add. peformance controls	Yes	Yes	Yes	Yes	Yes
World rank controls	Yes	Yes	Yes	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
JGP mean	1.840	1.840	1.115	1.115	1.044
Observations	16821	16764	16821	16764	12119
R^2	0.141	0.301	0.551	0.615	0.615

Nationalistic bias

Significant advantage when there is a compatriot judge

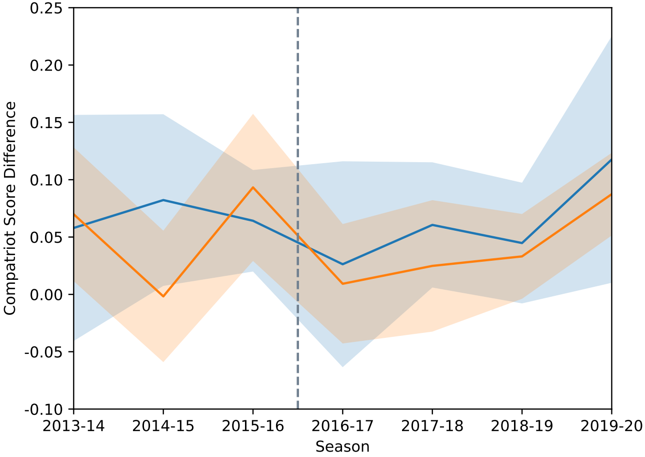
Table 2: Estimated nationalistic bias in the full sample

	Artistic score (std.)			Technical score (std.)		
	(1)	(2)	(3)	(4)	(5)	(6)
Compatriot	0.066***	0.046***	0.049***	0.044***	0.014**	0.020***
	(0.010)	(0.009)	(0.008)	(0.014)	(0.007)	(0.007)
Performance controls	Yes	Yes	Yes	Yes	Yes	Yes
World rank controls	–	Yes	Yes	–	Yes	Yes
Skater × Season FEs	–	–	Yes	–	–	Yes
Skater FEs	Yes	Yes	–	Yes	Yes	–
Round FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	16764	16764	16589	16764	16764	16589
R^2	0.867	0.891	0.937	0.708	0.911	0.933

Standard errors clustered at the event level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

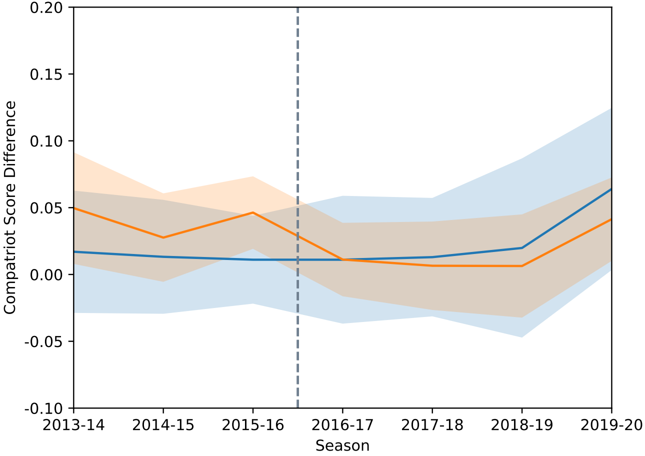
No reduction in nationalistic bias due to the reform

Figure 3: Nationality bias in the artistic score



No reduction in nationalistic bias due to the reform

Figure 4: Nationality bias in the technical score



No reduction in nationalistic bias

Table 3: Estimated effect of transparency on nationalistic bias

	Artistic score (std.)		Technical score (std.)	
	(1)	(2)	(3)	(4)
Compatriot	0.070***	0.035*	0.038***	0.032**
	(0.019)	(0.019)	(0.012)	(0.012)
Compatriot × Non-JGP	-0.006	0.018	-0.032*	-0.022
	(0.026)	(0.030)	(0.017)	(0.018)
Compatriot × Post	-0.042*	0.001	-0.035**	-0.024
	(0.024)	(0.023)	(0.015)	(0.018)
Compatriot × Post × Non-JGP	0.040	0.014	0.049**	0.046*
	(0.036)	(0.036)	(0.024)	(0.025)
Add. performance controls	Yes	Yes	Yes	Yes
Skater × Season FEs	–	Yes	–	Yes
Skater FEs	Yes	–	Yes	–
World rank controls	Yes	Yes	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes
Observations	16764	16589	16764	16589
R^2	0.884	0.937	0.911	0.933

Standard errors clustered at the event level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Additional Results

1. Larger decrease in score dispersion when there is greater public attention:
    - proxy public attention using average skater rank in the round
    - speaks for reputation concerns as driver

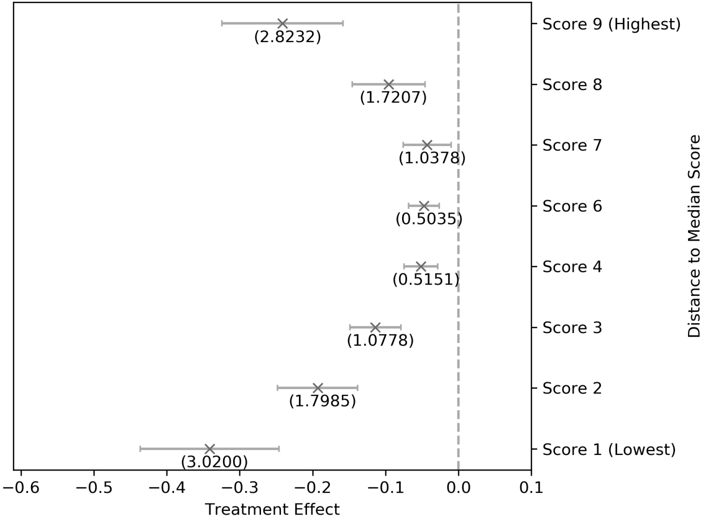
Additional Results

1. Larger decrease in score dispersion when there is greater public attention:
   - proxy public attention using average skater rank in the round
   - speaks for reputation concerns as driver
2. Post-reform, judges award more similar subscores for different components (higher "consistency"):
   - proxy for accuracy: correlates positively with judge experience, closeness to median score, and use of non-integer scores
   - suggestive evidence for increase in effort
3. No evidence for sequential learning about fellow judges
   - conformity effect does not increase with time in the same panel
4. No evidence for changes in judge selection

Thank you very much!

Backup slides

Changes in deviation of individual judges in panel



Effect is larger for more prestigious rounds

Table 4: Heterogeneous effects by average rank of skaters in the round

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	-0.001	-0.006	0.014	-0.025	-0.027
	(0.038)	(0.041)	(0.021)	(0.025)	(0.024)
Post × Non-JGP	-0.119***	-0.140***	-0.024	-0.032	-0.015
	(0.043)	(0.046)	(0.028)	(0.030)	(0.032)
Round quality × Non-JGP	0.071***	0.063***	0.000	-0.012	-0.016
	(0.015)	(0.017)	(0.012)	(0.014)	(0.015)
Round quality × Non-JGP × Post	-0.080***	-0.087***	0.018	0.008	-0.009
	(0.021)	(0.025)	(0.015)	(0.017)	(0.018)
Skater FEs	—	Yes	—	Yes	Yes
Additional performance controls	Yes	Yes	Yes	Yes	Yes
World rank controls	Yes	Yes	Yes	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.142	0.301	0.550	0.615	0.615

Standard errors clustered at the event level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Similar effects on subscore consistency as proxy for effort

Table 5: Effect of transparency on within-judge consistency of scores

	SD of artistic subscores		SD of technical subscores		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.017***	0.012***	0.021	-0.027*	-0.026*
	(0.004)	(0.004)	(0.014)	(0.014)	(0.015)
Post × Non-JGP	-0.016***	-0.017***	0.005	-0.007	0.009
	(0.005)	(0.004)	(0.018)	(0.016)	(0.016)
Add. performance controls	Yes	Yes	Yes	Yes	Yes
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
JGP mean	0.219	0.219	1.034	1.034	1.051
Observations	150458	150458	150431	150431	108675
R^2	0.041	0.090	0.233	0.360	0.342

Standard errors clustered at the event level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Limited heterogeneity by presence of compatriot judge

	SD of artistic subscores		SD of technical subscores		
	(1)	(2)	(3)	(4)	(5)
Compatriot	0.019	0.018	0.026**	0.017	0.014
	(0.027)	(0.031)	(0.011)	(0.017)	(0.017)
Compatriot × Non-JGP	0.066*	0.066*	0.010	0.026	0.023
	(0.036)	(0.038)	(0.015)	(0.022)	(0.022)
Compatriot × Post	-0.005	0.029	0.005	0.017	0.007
	(0.034)	(0.040)	(0.014)	(0.020)	(0.021)
Compatriot × Post × Non-JGP	-0.042	-0.087*		-0.022	-0.010
	(0.047)	(0.049)		(0.030)	(0.033)
Add. performance controls	Yes	Yes	Yes	Yes	Yes
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.315	0.448	0.641	0.693	0.690

Standard errors clustered at the event level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

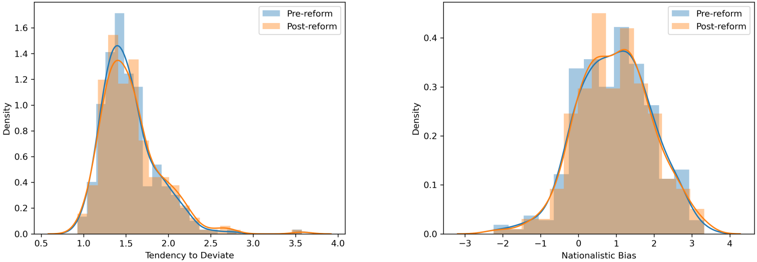
No evidence for conformity through social learning

	SD of Artistic Score		SD of Technical Score	
	(1)	(2)	(3)	(4)
Starting number	0.001	-0.001	0.000	0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Starting number × Post	-0.003	-0.001	0.001	0.001
	(0.003)	(0.002)	(0.002)	(0.002)
Starting number × Non-JGP	-0.019***	-0.015***	-0.002	-0.000
	(0.006)	(0.005)	(0.003)	(0.004)
Starting number × Non-JGP × Post	0.020**	0.015**	0.005	0.003
	(0.008)	(0.007)	(0.005)	(0.005)
Skater FEs	—	Yes	—	Yes
Add. performance controls	Yes	Yes	Yes	Yes
Skating group FEs	Yes	Yes	Yes	Yes
Observations	12861	12788	12861	12788
R^2	0.412	0.552	0.739	0.787

Standard errors clustered at the event level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

No evidence for changes in judge selection

Figure 5: Distribution of baseline judge-level scoring proxies



(a) Score accuracy proxy by judge

(b) Nationalistic bias proxy by judge