Cultural Bias: Gymnasts, Judges, and Bilateral Trade Agreements

Brian P Callahan
Seton Hall University

Sean E. Mulholland
Western Carolina University

Kurt W. Rotthoff*
Seton Hall University

Abstract

Bilateral trade agreements may arise from and enhance shared cultural norms across countries. Greater cultural understanding or similarities may be revealed through a number of avenues. We investigate one particular avenue: the judging of gymnastic performances. Using execution and difficulty scores from the 2009 World Gymnastics Championships, we find that gymnasts realize a greater return in terms of execution score for each bump in difficulty when the home countries of the judge and the athlete share a bilateral trade agreement.

*Kurt W Rotthoff is Associate Professor of Economics and Finance in the Stillman School of Business at Seton Hall University, JH 674, 400 South Orange Ave. South Orange, NJ 07079, 973.761.9102, Rotthoff@gmail.com. Sean Mulholland is Professor of Economics at Western Carolina University, 224D Forsyth, Cullowhee, NC 28723 828.227.3169, semulholland@wcu.edu. Brian Callahan is a graduate of Seton Hall University. A special thanks to Jack Harrington and the participants at the 2011 Eastern Economic Association meeting for helpful comments. Brian Callahan was an undergraduate at Seton Hall University when working on this project; he received a research grant for this project as part of a course receiving a grant from the Charles G. Koch Charitable Foundation. Errors or deficiencies that have survived are most assuredly ours alone.

I. Introduction

In David Ricardo's seminal work, "On the Principles of Political Economy and Taxation" (1817), he reveals the welfare gains that can be realized through specialization and exchange. These benefits are usually modeled in terms of lower prices and greater production for those directly involved with the consumption and production of the goods or services. However, the exchange of goods and services may be the result of or serve as a conduit for greater understanding and trust between those willing to voluntarily trade.

In this sense, trade and sentiment is a two-way street. In one direction trust and sentiment affects trade: Gupta and Yu (2007) find that government action and public sentiment impacts the levels of economic activity between two countries. Guiso, Sapienza, and Zingales (2009) find that language and ethnicity play an important role as well. They also find that similar countries, and countries that have high levels of trust, tend to have high capital flows between them. Once trade does take place, it can positively affect sentiment (Schiff and Winters, 1998).

This study seeks to determine whether the benefit of these trade agreements or the characteristics that lead to trade agreements extend beyond the traditional welfare measures. Using a unique dataset from the 2009 World Gymnastics' Championship we analyze whether the greater cultural understanding, as measured through trade agreements, implicitly affects the judging of gymnastic performances.

We find that gymnasts who are scored by judges from trading partner countries realize a greater return in terms of execution score for each increase in difficulty than do gymnasts who are evaluated by judges from non-trader partner countries. Shared cultural norms, such as a shared religion or language, or a reasonable understanding of different cultural norms may partially explain why countries are trading partners. Once cross-country trade is expanded, the exchange of goods and services may serve as a conduit for ideas that enable greater cross-cultural understanding. Cultural norms and views on beauty and expression may be better understood or even shared. This may affect how judges assess athletic and artistic expression. In gymnastics, a judge's view of what is and what is not an execution error, particularly on new or relatively novel elements or combinations, may be more closely aligned with athletes who share the judge's cultural framework.

Finding a cultural relationship bias has implications far beyond sports competition. Cultural bias might also affect how employers assess job applicants, colleagues assess each other's productivity, and whether negotiators are able to come to a contractual agreement. Using trade as a proxy for sentiment, we provide evidence that the positive sentiment towards one another extends beyond traditional welfare measures. The next section presents the data. The third section discusses our methodology. Section four presents our results and the final section concludes.

II. Data

To determine whether the cultural familiarity between a judge and an athlete results in higher scores, we match each athlete-judge country pair from the 2009 World (Artistic) Gymnastic Championships, to these respective countries' bilateral trade agreement(s). We use performance data from this particular championship because its

unique format reduces the selection bias problems present in most other gymnastics events (see Damisch, Mussweiler, and Plessner, 2006).[1]

First, there is no team competition. Most elite level international gymnastics meets have a team competition in which teams are allotted performance slots. Coaches then strategically place their athletes to maximize the team score. This traditionally means ordering the athletes from the lowest expected score to the highest. Without the team competition, coaches are not seeking to maximize the team score. Therefore, our data are less likely to suffer from bias due to strategic sorting.

Second, in the preliminary round, countries are randomly assigned one to three starting spots to place their athletes. These spots are then assigned to each athlete, by the country's governing body, within a given session, event, and order in that event. Judges therefore have the opportunity to measure an athlete's performance relative to the other athletes based on the overall performance order during the entire competition, the order in which they appear in a given session, and at the smallest level, the order in which they appear in a given rotation.

In women's gymnastics there are four different events (vault, uneven bars, beam, and floor) while the men participate in six events (vault, floor, pommel horse, rings, high bar, and parallel bars). The structure of the competition allows for enough recovery time between events, so that the athlete's performance on each event should, barring injury, be independent. Based on their performance in the preliminary round, athletes can make the finals in the individual all-around competition or for one or more individual events. The finals for each apparatus *is* structured in the traditional gymnastics way: in order from the lowest scoring finalist to the highest. For this reason, we only use data from the preliminary rounds. For each of the ten events, we observe between 106 and 134 performances; the number varies based upon the number of athletes attempting to make the finals in either the all-around or on a specific apparatus.

In addition to the unbiased ordering, the FIG (Federation Internationale de Gymnastique), the gymnastics governing body, completely overhauled the scoring system for elite level gymnastics in 2006.[2] Under this new system, scores are now determined by two separate panels of judges evaluating two components of the routine: difficulty and execution. The difficulty of a routine is initially determined by each element and combinations of elements that are planned for the routine. A panel of judges then evaluates the routine as it is actually performed and (rarely) adds or (usually) subtracts points for changes in the routine (such as under-rotation) or penalties (primarily given for athletes stepping out of bounds), to determine the final difficulty score. The difficulty score is theoretically infinite and is determined by the athlete when they design and perform their routine, meaning it is, in theory, exogenous to the execution judges. Furthermore, much of the difficulty score is objectively set by predetermined values for each element and set of combinations. Therefore interpretation and by extension cultural biases likely play little to no role in determining the difficulty score.

The execution score evaluates how well the athlete performs a given routine. Each athlete starts with an execution score value of 10. Although deductions from the

[1] For instance, Damisch, Mussweiler, and Plessner (2006) find a sequential order bias; that an athlete's score is influenced by the athlete who performed immediately before them. Rotthoff (2013) finds that this sequential order bias does not exist at this competition.

[2] This change came after a judging controversy in the 2004 Athens Olympics.

execution score can reflect purely technical errors, many other possible deductions come from the artistry of the routine, such as poor rhythm, additional hops or swings, or incorrect body position. Given that the execution score is determined by the judge's perception of how well the routine was executed, the execution is quite subjective and thus more likely to reflect a judge's cultural and artistic biases.

The execution and difficulty scores are then added together at the end of each routine for an overall score. The overall score is finalized before the next contestant makes their attempt. The average and standard deviation of scores for women are shown in Table 1. The summary statistics for men are shown in Table 2.[3]

Table 1 – Summary statistics for the women's events.

Summary Statistics (women)

Variable	Vault	Uneven Bars	Balance Beam	Floor
Participants	107	113	118	113
Mean Difficulty Score	4.94	4.89	4.99	4.92
Standard Deviation of Difficulty Score	0.706	1.194	0.650	0.564
Mean Execution Score	8.24	6.91	7.21	7.37
Standard Deviation of Execution Score	0.904	1.517	1.161	0.778

Table 2 – Summary statistics for the men's events.

Summary Statistics (men)

Variable	Parallel Bars	High Bar	Rings	Floor	Vault	Pommel Horse
Participants	127	127	126	134	122	132
Mean Difficulty Score	5.31	5.31	5.43	5.51	5.31	5.14
Standard Deviation of Difficulty Score	0.88	1.00	0.91	0.79	0.88	0.90
Mean Execution Score	8.07	7.80	7.94	8.16	8.07	7.68
Standard Deviation of Execution Score	0.78	0.85	0.66	0.96	0.78	1.17

The 2009 World (Artistic) Gymnastic Championships meet is the first elite level competition without a team competition to use the new, overhauled scoring system. In addition to measuring potential cultural bias, we need to control for other forms of bias that may be present given the characteristics of gymnastic competition. Flôres and Ginsburgh (1996), Bruine de Bruin (2005), and Page and Page (2010) all find that when a participant performs affects their score. Flôres and Ginsburgh (1996) find the day an artist competes impacts that artist's final standing. While Bruine de Bruin (2005) and Page and Page (2010) find that order biases in the results of the "Eurovision" and "Idol" song contests, respectively. Following the literature, we control for performance order by utilizing the athletes' order in the overall competition.

[3] Given the differences in means and standard deviations, it is necessary to use a relative scoreto aggregate the data across different events.

Although athletes are randomly assigned a performance slot, performance bias might also be driven by a few very talented people.[4] This might come from a trend setting athlete or team of athletes attempting to change the culture associated with artistic expression or innovative elements. To address this possibility, and because ability is very difficult to measure on its own, we control for athletes who come from superstar countries (following Morgan and Rotthoff, 2014, and Rotthoff, 2015). The superstar countries are shown in Table 3, for women, and Table 4, for men. To be defined as a superstar country, the athlete's country has to have won at least three medals, in a given event, in one of the World's competitions (2001-2003 and 2005-2007) or the Olympics (2000, 2004, and 2008).

Table 3 – Countries that are considered 'superstar' countries for women's events.

Super Star Countries (women)			
Vault	Uneven Bars	Balance Beam	Floor
USA	USA	USA	USA
Russia	Russia	Russia	Romania
China	China	Romania	
Germany		China	

Table 4 – Countries that are considered 'superstar' countries for men's events.

Super Star Countries (men)					
Parallel Bars	High Bar	Rings	Floor	Vault	Pommel Horse
China	Germany	China	Canada	China	China
S. Korea	Slovakia	Bulgaria	Romania	Romania	Romania
		Italy		Poland	Japan

Each event has two panels of judges: one panel calculates the difficulty score and one panel assesses the execution of each routine. Judges may only serve on one panel for an event. Using information from GymnasticsResults.com, we observe the home country of each judge on each execution panel. These are reported in Tables 5 and 6. We do not have the country of each judge on the difficulty panel.

Table 5 – Country of the execution judges, by event.

Country of Execution Judges (women)			
Vault	Uneven Bars	Balance Beam	Floor
Mexico	N. Korea	India	Slovenia
Bulgaria	Egypt	Ireland	Germany
S. Korea	Norway	Portugal	Venezuela
Italy	Canada	Argentina	Lithuania
Romania	Brazil	France	China
Ukraine	Germany	Israel	Russia

[4] It is possible that the most talented individuals capture a different level of bias than other athletes in the competition. Controlling for these superstar countries allows for those at the highest level of talent to have a different impact from these biases.

Table 6 – Country of the judges, by event.

Country of Execution Judges (men)

Parallel Bars	High Bar	Rings	Floor	Vault	Pommel Horse
Netherland	Algeria	Bulgaria	Japan	Mexico	Slovenia
S. Korea	Portugal	France	Venezuela	New Zealand	Russia
Lithuania	Austria	Germany	Luxemburg	Belarus	Portugal
Argentina	Ukraine	Qatar	Romania	Germany	Brazil
Czech Republic	Hungry	Jordan	Egypt	Canada	N. Korea
Poland	Great Britain	South Africa	Italy	Israel	Denmark

Not only do we know the country of each judge on the execution panel, we also have the exact score submitted by each judge for each routine. This means that we have six scores for each athlete for each event. In gymnastics the highest and lowest scores are dropped, and then the overall execution score is the average of the remaining scores. However, because we are interested in each judge's interpretation of the routine, we use all six judging scores for this analysis. Because we are interested in the cultural bias that may be present across countries, we exclude all athlete/judge pairs that are from the same country. Although it is likely that athletes and judges share a cultural understanding, this understanding is likely to come from more personal experience. Moreover, coding trade within a country as similar to those with bilateral trade agreements across counties may conflate our results.

Given we know the home country of each execution judge, the score given by each of these individual judges, and the country represented by each athlete, we are able to extend the literature on assessment bias: Goldin and Rouse (2000) and Page and Page (2010) find that the order of competition matters, Price and Wolfers (2010) find a racial bias in basketball refereeing, Zitzewitz (2006 and 2014) finds a nationalism bias in figure skating, and Morgan and Rotthoff (2014) find a difficulty bias in gymnastics. We exploit this information to measure whether athletes receive higher execution scores when the home countries of the judge and the athlete share a bilateral trade agreement.

We define our bilateral trade agreement as a formal arrangement between two countries granting one another preferred trading status in some area of their respective economies. There are three ways in which we classify countries as bi-lateral trading partners. First, countries can have formal agreements directly with one another. For instance, the United States-Israel Free Trade Area Agreement (FTA) signed in 1985, was created to lower tariffs between the United States and Israel.

The second way is for two countries to be partners in a trade association. The largest and most publicized of these organizations is the European Union (EU). Other such organizations and agreements relevant to our calculations include European Free-Trade Agreement (EFTA), European Economic Area (EEA), NAFTA (North American Free-Trade Agreement), APEC (Asia – Pacific Economic Cooperation), MERCOSUR (Mercado Común del Sur; English – Southern Common Market), Group of 3, and ASEAN (Association of Southeastern Asian Nations).

The third way to be classified as sharing a trade agreement is for a country's trade organization to have a collective bi-lateral agreement with a country outside of that trade agreement. For example, the EU has a bi-lateral trade agreement with Argentina even though Argentina is not part of the EU, nor does it have trade agreements with individual

members of the EU. We match these trade agreement data with our gymnastics data, by matching the country of the athlete and the country of judge, to find those that share a bilateral trade agreement.[5]

III. Methodology

We have the execution score given by each individual judge, *j*, for each event, *v*, for each male and female gymnast's routine. To account for differences across each event, we model the judge's score relative to the overall execution score for athlete *i* in event *v* in equation 1.

$$\frac{JudgesExecutionScore_{ivj}}{OverallExecutionScore_{iv}} - 1 = RelativeExecutionScore_{ivj} \quad (1)$$

Thus if a particular judge's score, *Judges Execution Score*$_{ivj}$, is the same as the averaged score given to that athlete, *Overall Execution Score*$_{iv}$, the relative execution score, *Relative Execution Score*$_{ivj}$, is zero. This relative measure of each judge's execution score, relative to the overall score given to each athlete, allows us to measure a particular judge's score relative to the other judges' scores and allows us to aggregate the judges within each event and across events.[6]

A value greater than zero reveals that an execution judge viewed the athlete's routine more favorably that his or her peers. Although this may be due to a number of factors, one may stem from a cultural interpretation of the routine. The summary statistics, including both the relative and raw execution and difficulty scores are in Table 7.

Table 7: Summary Statistics

	Mean	Standard Deviation	Max	Min
Overall Score	13.12	1.542	16.8	6.725
Raw Execution Score	7.849	0.997	9.7	1
Relative Execution Score	-0.001	0.038	0.485	-0.889
Raw Difficulty Score	5.298	0.83	7.2	1.7
Order	62.65	36.24	135	1
Superstar	0.062	0.242	1	0
Cultural Relationship (Yes=1)	0.644	0.479	1	0
N	4518			

To control for the known biases in the literature we estimate equation 2:

[5] Do to inaccurate trade data we drop athletes from Kazakhstan and Kuwait.

[6] Given that there are different means and standard deviations across events, this method allows us to aggregate across events (as in Morgan and Rotthoff, 2013, who use a normalization process).

$$\begin{aligned}\text{Relative Execution Score}_{ivj} &= \beta_0 + \beta_1 \text{Overall Order}_{iv} + \beta_2 \text{Overall Order}_{iv}^{\,2} + \beta_3 \text{Diffculty Score}_{iv} \\ &+ \beta_4 \text{Super Star}_{vj} + \beta_5 E_v + \beta_6 \text{Male}_i + \beta_7 \text{Cultural Relationship}_{ij} \\ &+ \beta_8 (\text{Cultural Relationship} * \text{Difficulty})_{ivj} + \varepsilon_{ivj} \qquad (2)\end{aligned}$$

where the *Relative Execution Score* submitted by judge *j*, for athlete *i*, in event *v*, is a function of the athletes performance slot, *Overall Order* and *Overall Order squared*, the *Difficulty Score* for the athlete's routine, the superstar effect, *Super Star*, a vector of event specific dummy variables that control for any fundamental differences between the different events, *E*, a control for any fundamental differences in the male and female athletes, *Male*, and the *Cultural Relationship*, for which trade agreements are the proxy.[7]

Morgan and Rotthoff (2014) find a difficulty bias. This means that attempting a more difficult routine raises a gymnast's execution score, even though a) the scoring is theoretically independent and b) more difficult routines increase the probability of technical and artistic deductions. The presences of difficulty bias means that although cultural biases may directly affect the execution score, it is possible that routines containing more difficult elements or combinations may be subject to more cultural bias. One reason for this may come from the fact that more difficult routines likely include new elements and combinations that execution judges rarely witness. Without the benefit of well-defined technical or artistic standards, execution judges must use greater discretion when assessing a routine.[8] For this reason we run the model both with and without an interaction term of the cultural relationships and difficulty score.

IV. Results

If bilateral trade agreements enhance or measure cultural understanding, then, the effects can be measured through the coefficient on cultural relationships and, possibly, the interaction between cultural relationships and difficulty. A positive relationship would suggest that an athlete whose country has a trade agreement with the judge's home country realizes a benefit from a common cultural understanding or overall positive sentiment. Using a relative execution score we measure how each execution judge's score is relative to the average judge's score in Table 8. The first column excludes the interaction on cultural relationship and difficulty score. The second columns include this interaction. The Hausman test values of 3.12 for column one and 3.86 for column two reveal that coefficients estimated by the efficient random effects estimator are not statistically different than the ones estimated by the consistent fixed effects estimator. Therefore, both columns report the random effects estimator (Hausman, 1978).

[7] Because we are interested in whether a specific judges give higher score to athletes from country's with which they share a cultural understanding, we also, as a robustness check, estimate this equation with athlete fixed effects, to separate out a judges possible preference for a specific athlete – as opposed to all athletes from a specific country – and judge fixed effects to account for the possibility that specific judges may consistently give higher or lower scores for all performances. The results are qualitatively similar. We have also interacted *Male* and *E*. We continue to find similar results.

[8] Athletes might alter the difficulty level of their routines if the athlete or coach believes a judge will view their execution favorably. We do not have the ability to separate these effects out, thus we treat them as the same.

Without the interaction, the coefficient on cultural relationship in column 1 of Table 8 is insignificant. This implies that having a cultural relationship has no impact on an individual judge's execution score relative to the average judge's execution score.

We also find that more difficult routines are not statistically associated with a higher relative execution score. Thus judges appear to similarly account for difficulty score when calculating the athlete's execution score. This does not refute Morgan and Rotthoff (2014) findings that execution scores increase with difficulty; it only shows that on average all judges' execution scores are higher when an athlete attempts a more difficult routine.

Table 8: Results of equation 2. Finding that cultural relationships have a positive impact on the difficulty bias found in Morgan and Rotthoff (2014).

VARIABLES	(1) Relative Execution	(2) Relative Execution
Order	0.00009* (0.091)	0.00009* (0.087)
Order Squared	-0.00000 (0.131)	-0.00000 (0.126)
Difficulty Score	0.00044 (0.359)	-0.00109 (0.207)
Superstar	-0.00042 (0.532)	-0.00048 (0.468)
male	-0.00057 (0.589)	-0.00058 (0.585)
Cultural Relationship (Yes=1)	-0.00009 (0.939)	-0.01306* (0.062)
Cultural Relationship x Difficulty Score		0.00246** (0.044)
Constant	-0.00473 (0.119)	0.00324 (0.511)
Observations	4,518	4,518
R-squared	0.0008	0.0016
Event Specific FE	Yes	Yes
Athlete RE	Yes	Yes
Number of id	322	322
Overall Cultural Effect		-0.01060
Standard Error		0.00070
Overall Difficulty Effect		0.00137
Standard Error		0.00579

Robust pval in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In column 2 of Table 8, we present the results including the interaction of cultural relationship and difficulty score. The estimation reveals a positive and significant coefficient on the cultural relationship/difficulty score interaction term. Therefore athletes attempting more difficult routines receive an even greater increase in their execution score from judges with a cultural relationship than from judges without a cultural relationship.

This suggests that, although all judges appear to give higher execution scores for more difficult routines, one source of the difficulty bias found by Morgan and Rotthoff (2014) may be the shared cultural experience of a judge and gymnast. In terms of points this means that a gymnast attempting a one standard deviation more difficult routine, or an increase of 0.83 points from 5.3 to 6.13, receives an execution score that is 0.0536 points greater from a judge with a cultural relationship than from a judge without this relationship. The magnitude of the estimated cultural bias has the ability to alter the outcome of this event. For reference, a 0.0536 point increase on each event would be an increase of 0.214 in a woman's overall score. This is a large enough increase in score to move the second place gymnast, who was from Romania, into first place. On the uneven bars this also means that the third place finisher, from the People's Republic of Korea (North Korea), would have finished second if she shared a greater cultural understanding with the judges. Although the R-squared measures are small for the overall impact on the athlete's score, the existence of the bias itself, and the magnitude of the bias, are still relevant.

Our results suggest that judges deduct fewer execution points from more difficult routines if the gymnast is from a country with a cultural relationship. Gymnasts attempting difficult routines may be performing new elements or combinations whose technical and artistic standards may not be well-defined. Thus the execution score relies less on the accepted standards and more on a judge's interpretation of the artistry, form, body position, etc. Without well-defined standards, a judge's assessment maybe more susceptible to cultural interpretation. If the athlete's approach to these new elements or combinations is more similar to the judge's, then an athlete will receive a higher execution score for these more difficult routines.

These findings support the idea that cultural relationships extend beyond the international trade arena. Knowing a judge on the panel has a cultural relationship means that the athletes can expect a larger difficulty bias from this judge. Given the dependent variable is the judge's score relative to the average execution score, which includes that judges score, these estimates are underestimating the true effect of this finding.

V. Conclusion

Formal bilateral trade agreements increase the gains from trade. This exchange between nations is both reflective of and serves as a conduit for greater understanding and trust between citizens in each country. We use data from execution judges of the 2009 World (Artistic) Gymnastic Championships, finding that gymnasts who are scored by judges with a cultural relationship, using trading partner countries as a proxy, receive a higher difficulty bias and thus higher execution scores than athletes without a cultural relationship performing similarly difficulty routines. We find that for each one standard deviation increase in difficulty, a judge enters an execution score that is 0.0536 higher for athletes with a cultural relationship than those athletes without a cultural relationship.

This occurs even though the execution judges are not charged with determining the difficulty score. These results suggest that the benefit of these cultural relationships or the characteristics that lead to trade agreements extend beyond the traditional welfare measures.

These findings shed light on one possible source of the difficulty bias found in Morgan and Rotthoff (2014). Our results also support the suggestion by Glejser and Heyndels (2001) that accurate comparisons of different people, in their case musical auditions, requires music that is the same level difficulty. Finding a non-welfare enhancing bias in the judging process results in an inefficient, or at least suboptimal, outcome.

If our hypothesis is correct and cultural differences measured through trade agreements are associated with a larger difficulty bias in execution scores, a gymnast might want to maximize the more objective score (difficulty) than the less objective (execution). The size of this effect is large enough to alter the overall standings and, in some cases, the composition and order of the top three athletes on the podium.

Works Cited

Burgess, N. and G. Hitch, (1999). Memory for serial order: a network model of the phonological loop and its timing *Psychological Review* 106, 551–581

Damisch, L, T. Mussweiler, and H. Plessner, (2006). Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments *Journal of Experimental Psychology Applied* 12, 166

Goldin, Claudia and Cecilia Rouse, (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians *The American Economic Review,* Vol. 90(4): 715-741.

Gershberg, F. and A. Shimamura, (1994). Serial position effects in implicit and explicit tests of memory *Journal of Experimental Psychology: Learning, Memory and Cognition* 20, 1370–1378

Glejser, H. and B.Heyndels, (2001). Efficiency and inefficiency in the ranking in competitions: the case of the Queen Elisabeth music contest *Journal of Cultural Economics* 25, 109–129.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, (2009). Cultural Biases in Economic Exchange *The Quarterly Journal of Economics*. 124(3): 1095-1131.

Gupta, Nandini and Xiaoyun Yu. (2007). Does Money Follow the Flag? Working Paper. Available at SSRN: http://ssrn.com/abstract=1316364

Hausman, Jerry A. (1978). Specification tests in econometrics *Econometrica* 46: 1251-1271.

Morgan, Hillary N. and Rotthoff, Kurt W. (2014). The Harder the Task, the Higher the Score: Findings of a Difficulty Bias *forthcoming in Economic Inquiry*

Page, Lionel and Katie Page. (2010). Last shall be first: A field study of biases in sequential performance evaluation on the Idol series. *Journal of Economic Behavior & Organization* 73(2): 186–198.

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

Ricardo, David. (1817). *On The Principles of Political Economy and Taxation* London: John Murray, Albemarle-Street.

Rotthoff, Kurt W. (2015) "(Not Finding a) Sequential Order Bias in Elite Level Gymnastics" *Southern Economic Journal* Volume 81, Issue 3, pages 724–741

Schiff, M. and L.A. Winters. (1998). Dynamic and Politics in Regional Integration Arrangements: An Introduction. *The World Bank Economic Review*. 12(2): 177-95.

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

Zitzewitz, Eric, (2014). Does Transparency Really Increase Corruption? Evidence from the 'Reform' of Figure Skating Judging. *Journal of Sports Economics* 15: 3-30