Points Systems in Judged Performance Sports: Case Study of the Figure Skating International Judging System

By
Patrick Costley

Senior Honors Thesis
Economics
University of North Carolina at Chapel Hill

April, 2020

Approved:

_________________________
Dr. Kyle Woodward, Thesis Advisor

_________________________
Dr. Jane Fruehwirth, Reader
## Contents

1. Acknowledgements  
2. Abstract  
3. Introduction  
4. Background  
5. Related Literature  
6. Theoretical Model  
7. Theoretical Results  
8. Empirical Model  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>1</td>
</tr>
<tr>
<td>Abstract</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>Background</td>
<td>7</td>
</tr>
<tr>
<td>Related Literature</td>
<td>9</td>
</tr>
<tr>
<td>Contracts</td>
<td>9</td>
</tr>
<tr>
<td>Figure Skating Judging</td>
<td>11</td>
</tr>
<tr>
<td>Theoretical Model</td>
<td>13</td>
</tr>
<tr>
<td>Introduction to the Theoretical Model</td>
<td>13</td>
</tr>
<tr>
<td>Full Information Points-Maximization</td>
<td>16</td>
</tr>
<tr>
<td>Incomplete Information Points-Maximization</td>
<td>18</td>
</tr>
<tr>
<td>Competitive Agents</td>
<td>21</td>
</tr>
<tr>
<td>Theoretical Results</td>
<td>23</td>
</tr>
<tr>
<td>Desirability</td>
<td>23</td>
</tr>
<tr>
<td>Difficulty</td>
<td>24</td>
</tr>
<tr>
<td>Applications</td>
<td>25</td>
</tr>
<tr>
<td>Empirical Model</td>
<td>27</td>
</tr>
<tr>
<td>Relation to the Theoretical Model</td>
<td>27</td>
</tr>
</tbody>
</table>
1 Acknowledgements

This project is something I have been excited to pursue for several years; it is the convergence of several of my passions. Thank you to my advisor Dr. Kyle Woodward for diving head-first with me into new and challenging territory and helping me to bring this idea to fruition. I want to offer special thanks to Dr. Woodward and Dr. Jane Fruehwirth for the level of personal support you provided me beyond the duties of faculty; it means a lot to me and I hope you understand the ways you have helped me as I end my undergraduate career and prepare to move on in my career.

2 Abstract

This paper treats judged performance sports as a principal-agent contract. I model the factors that influence optimal points systems for judged performance sports such as figure skating. This model yields two distinct effects. The desirability effect indicates that elements that are more beneficial for the sporting federation’s output of interest should be given more points. The difficulty effect indicates that more difficult skills should be given fewer points. These effects often work in opposite directions; difficult elements are often desirable. Thus, federations must fully consider both effects to select an optimal system. These results suggest that if a points system is suboptimally focused on desirability, it could be improved by reducing the points allocated to difficult elements. It could also be improved by giving more points to more desirable elements or conversely increasing the deductions applied to undesirable elements.

Secondly, I analyze figure skating performances before and after a change from a ranking...
system to an additive points system to determine the effect of the change on performances. This analysis suggests the current figure skating judging system may be suboptimal and could be improved following the prescriptions of my model.

3 Introduction

The sport of figure skating is organized by the International Skating Union (ISU). The ISU hosts competitions in which skaters from around the world compete. Assuming skaters and the ISU are rational actors with individual interests, their behavior can be understood through economic modeling.

Rational skaters desire to win competitions. This is a corollary of the commonly understood assumption that individuals maximize utility. Individuals can derive utility from consumption – money. But they can also derive utility from non-money sources. For example, in a labor-leisure model, workers gain utility from consumption and leisure time.

For some individuals, figure skating can be work for pay that allows them to derive utility from consumption. However, this study focuses on competitive skaters rather than professional skaters.

Competitive figure skaters are still rational utility maximizing actors. Christopher Dean and his ice dance partner Jane Torvill were 1984 Olympic figure skating champions. After the Olympics, they skated professionally for pay for ten years. The ISU changed the eligibility requirements to allow skaters who had skated professionally to compete. So Torvill and Dean returned to compete in the 1994 Olympics. In an interview he gave for UPI, Dean said, “the International Skating Union wants the best competitors at the premier event. It’s like the
tennis and basketball at the Summer Olympics. The difference is that as a professional I get paid. As an amateur I get paid in a different way” (Phillips, 1993). According to this reasoning, considering skaters as desiring to win competitions is a corollary of considering them as utility maximizers. There is some attribute that is unique to figure skaters that causes them to sort themselves into figure skating competitions and derive utility in a form different to money from competing.

The opportunity to be judged in comparison to competitors (and possibly win) is the defining characteristic that differentiates performing in competition from performing professionally for skaters. This can be inferred from Christopher Dean’s statement and also demonstrated in the ways skaters describe their rivalries. For example, in a Japan Times article about Yuzuru Hanyu’s defeat at the 2019 Grand Prix Final to rival Nathan Chen, Hanyu says, “I want to be better. Nathan makes me stronger and makes it more fun to skate” (Gallagher, 2019). This statement indicates that losing to other skaters creates a challenge to improve that motivates the skater and enhances the utility the skater derives. Winning against skilled rivals provides more utility. Thus, a rational competitive skater derives utility from placing well at competitions.

A rational ISU desires to maximize some output of interest. Conventionally, the output of interest for an organization would be profit. This may be true. Alternatively, the ISU might benevolently desire the success of figure skating as a sport. It is not necessary to precisely define what this output is. For the purposes of this paper I call it viewership because the number of people watching figure skating or their level of engagement with the sport is more or less a similar indicator to the profit the ISU might earn or the success of the sport the ISU might like to see.
Therefore, skaters and the ISU are rational actors with distinct interests. They interact with one another in pursuit of these interests. Their interactions are fundamentally defined by the system of judging skating performances that the ISU sets because the judging system determines the placement of skaters at competitions and the performances the skaters give determine viewership. The judging system the ISU currently uses is called the International Judging System. I explain the structure of this system and the reasons behind its conception in the Background section. Essentially, this is a points system and the skater who earns the greatest sum of points wins.

However, there is speculation as to whether this is really the best points system. For example, Tabb (2018) summarizes in his article for Quartz some of the concerns people have with the system. The system offers a lot of points for performing a quadruple jump, which is difficult and potentially dangerous. The judging system incentivizes skaters to perform these difficult jumps. But performing them carries a higher risk of injury than simpler jumps. Providing too much incentive to perform such difficult jumps risks forcing fan-favorite skaters to sit out competitions because of injury, which would be bad for viewership. Additionally, there is speculation as to how well the judging system achieves a desirable balance between the jumps and other aspects of performance. In other words, there is uncertainty as to whether the current judging system incentivizes skaters to give performances that are optimal for viewership.

In economics, the question of determining optimal interactions between skaters and the ISU can be characterized as a principal-agent problem within contract theory. The principal is the ISU. The agents are the skaters. The ISU selects the judging system, which is the contract. The skaters and the ISU each seek to optimize their interests according to the
contract. The question this paper seeks to answer is: how can the ISU select a judging system to award points to competitive skaters such that the optimal performance for skaters also maximizes viewership?

This paper does not seek to prescribe an ideal system. It instead seeks to understand the relationships between skaters and the ISU and how behavior is affected by the judging system, and it seeks to determine how the ISU might make changes to the system it uses so that it can more effectively pursue its interests subject to how skaters perform. I find two main theoretical results.

The first result is unsurprising: all else equal, performing a skill that is more desirable for viewership should be awarded more points. Conversely, performing a skill that is less desirable or undesirable for viewership should be given fewer points, or a larger deduction. I term this the desirability effect.

The second result is more surprising: all else equal, a skill that is more difficult for the skater to perform should be given fewer points, and a skill that is easier for a skater to perform should be given more points. This may seem counterintuitive because one might expect that performing something difficult should be rewarded with more points. But this result comes from the fundamental assumption that the ISU is viewership maximizing. Consider two elements that viewers like equally. One is easier for skaters than the other, so it can be performed in a way that is more beneficial for viewership with the same amount of effort. The ISU should incentivize skaters to perform the easier skill by giving it more points. I term this the difficulty effect.

In practice, skills that are more difficult are often more desirable for viewership. This means that the desirability effect and the difficulty effect are working in opposite directions.
The responsibility for determining which effect should be given greater consideration lies with the ISU.

Following my theoretical analysis, I discuss empirical results to determine if the existing judging system is optimal. This analysis suggests that the implementation of the International Judging System resulted in suboptimal performances. Analysis of time trends in performances does not yield statistically significant results. Nonetheless, the signs of the regression parameters are consistent with a judging system that is becoming less optimal. These results suggest the judging system could be improved by modifications to the points awarded according to the desirability and difficulty effects.

The analysis of this paper can be generalized to any judged performance sports. Sports competitions can take many structures. In some sports, two teams or individuals compete against one another and matches are played as a part of a tournament. In other sports, athletes can be metrically evaluated against one another, such as in races where athletes compete for the fastest times. Figure skating falls into a category in which athletes perform an action that cannot be metrically evaluated. Performance is instead evaluated by a panel of judges. Other judged performance sports include artistic and rhythmic gymnastics, diving, some skiing and snowboarding sports, and skateboarding.

I explore the contracts question of optimal judging systems for athlete and sporting federation interests through the lens of figure skating. But these results are relevant for any such event with this principal-agent relationship. These sports share a large combined viewer base, and sporting federations could stand to benefit their interests from implementing informed judging systems to incentivize optimal performances by athletes.
4 Background

The International Judging System (IJS) is relatively new. It was implemented in 2004 after a scandal at the 2002 Olympics that highlighted long-held concerns about bias in the 6.0 system (Reid, 2006). In the 6.0 judging system, judges awarded a score out of six to each skaters in the categories of technical merit and performance. These scores were used to rank skaters. The skater who had the majority of 1 votes would be awarded first place. The remaining 1 votes would become 2 votes and then the skater with the most 2 votes would be awarded second place, and so on to assign all ranks (SDFSC, n.d.). It was not the magnitude of the score that mattered, it was the order that was important. Where the 6.0 system was a ranking system, in the new IJS, skaters accrue points based on the elements they complete and the overall quality of their performance. As Clarey (2002) reports, in the 6.0 ranking system, judges were able to effectively trade votes between disciplines to help compatriot skaters succeed. The IJS is intended to be more objective because it evaluates skaters individually instead of in comparison to other skaters, and on the sum on individual components of a performance instead of on holistic impression.

Initially, in the IJS, the ISU did not report which judge gave which score. This was intended to disrupt the ability of officials to collude. Now, the ISU does report which judge gave scores, so there is accountability for judges to act fairly. There remain concerns about judging bias affecting competition results. As Abad-Santos (2014) describes, many people believe bias affected the outcome of the 2014 Olympics. While this paper must assume judges are able to represent the interests of the ISU, it may be perpetually difficult in judging to ensure judges do not act out of their own interests or out of nationalistic interest.
In the IJS, for each program, skaters receive a Technical Element Score (TES) and a Program Components Score (PCS). The program components are an evaluation of overall quality in the categories of skating skills, transition steps, performance and execution, interpretation of the music, and choreography. Skaters are scored from 0 to 10 in intervals of 0.25. Consequently, there is a maximum PCS. The TES is the score given for the moves the skater executes. Each element has a pre-assigned base value for the element’s difficulty, and judges add or subtract from this base value with their Grade of Execution score (GOE). The TES has no maximum, as skaters could potentially perform more and more difficult elements.

Joe Inman was one of the judges involved in creating the IJS. In a 2005 interview, not long after the initial implementation of the IJS, Inman explained how the TES and PCS were meant to make up the program. He said, “you get a certain percentage for the technical...and a certain percentage for the — it’s supposed to be 50-50, it’s not quite that, I understand. It’s supposed to be a little more in the technical, but very slightly” (TheSkatingLesson, 2019). This suggests that the ISU believed that a good figure skating performance should have balanced artistry and technical difficulty. Indeed, in the first years under the IJS, there was an artistic balance that was close to 50-50 as Inman mentioned. For example, in the 2006 Olympics, the men’s free skate champion, Evgeni Plushenko, was awarded a TES of 85.25 and a PCS of 82.42. However, in the 2018 Olympic, the men’s free skate winner, Nathan Chen, was awarded a TES of 127.64 and a PCS of 87.44. This suggests that the IJS has not yielded the type of program balance that the ISU believed would be optimal.

In summary, the challenge for the ISU is selecting a judging system that is fair to skaters and incentivizes desirably balanced performances.
5 Related Literature

In this section, I discuss how the research question can be understood in the context of contract theory and how this question differs from other questions of contracts. The topic is judging in sports, specifically figure skating. I discuss other research related to figure skating judging and ranking systems.

5.1 Contracts

When considering principal-agent interactions in the context of figure skating, the ISU is the principal and the skaters are the agents. The ISU selects a points system that determines how points will be distributed to skaters. This points system is the contract. Principal and agent interests are distinct. The ISU objective is maximizing some output I call viewership, and the skater objective is placing as high as they can in competition rankings, where rank is determined by points.

Harris and Raviv (1978) are concerned with incentive problems that may arise in principal-agent relationships as they may relate to employment contracts and other scenarios. The question of the paper is how to optimize contracts subject to these problems. In their model, the agent performs an action subject to random exogenous variables to produce an output. Subject to a contract, the principal and agent share the output. The agent has a negative utility of performing the action, the principal wants the agent to perform the action, and so must structure the contract to optimize their share of the output while the agent optimizes their own utility. They vary this model according to different levels of risk, levels of knowledge of exogenous random variables, and levels of risk aversion to optimize.
This case study of figure skating deals with a problem similar to the one Harris and Raviv study. However, there are some contextual differences. In employment contracts, the principal and the agent split the output. Any output that the agent enjoys is deducted from the principal. In the case of sporting judging systems, it costs principals nothing to award points to agents.

Another way that this sporting problem differs from employment contracts is the strategic interaction between agents. Where in employment problems, agent utility is a function of the payout, in sports judging problems, utility is a function of the payout in relation to the payout of other agents. So consider game theory approaches to strategic interactions. Kovenock and Roberson (2011), for example, study a Colonel Blotto game. Two players possessing $m$ identical units of resources distribute their resources over $n$ battlefields where the player with the most units on a battlefield wins the battle and the player who wins the most battles wins the game. Kovenock and Robeson’s paper models a particular version of this game with incomplete information. They determine an equilibrium strategy that each player should take.

This is fundamentally a question of multidimensional effort decisions. In the context of figure skating, this can be compared to the ways that skaters choose to allocate training resources such as time: skaters must choose how to allocate time to training various elements. This develops the notion of an employment contract further, because it contributes the potential of agent performing multiple actions to maximize points. Skaters have the ability to change how well they can perform different actions by where they allocate training resources. This is because in competitions, judges judge skaters by the sum of all aspects of a performance and not by an element by element comparison between skaters. However, this
study differs still from a Blotto game, because there are more than two agents. Furthermore, the Blotto game is more similar to the 6.0 judging system because the winner is determined by the agent who has the most battlefield wins – the most 1 votes. In contrast, the IJS additive points systems allows for cardinal comparisons between agents.

These studies are illustrative of the types of questions that appear in contract theory and game theory and the sorts of approaches to these questions. As I noted, the question of optimal judging systems is different from conventional principal-agent problems. It is related to the literature on contracts, but it is relevant to applications in judged performance sports, rather than in applications such as employment contracts. The multitude of judged performances sports constitutes a broad range of applications for this paper.

5.2 Figure Skating Judging

As stated in the Background section, a challenge the ISU faces is selecting a judging system that is fair to skaters. The 2002 Olympic figure skating judging scandal was an event in which long held concerns about judging systems manifested in an obvious and consequential way. Other research has focussed on the problems associated with judging sports.

Balinski and Laraki (2007) provide a useful summary of different approaches to judging systems and their associated problems. They differentiate ranking systems and measures or grades. In a ranking system, judges rank finite items in a set. A problem this approach presents is uncertainty as to the level of strategy judges use when ranking. For example, a judge might like skater A best and skater B second best. Desiring skater A to win, they rank skater B much lower than second. In a grading or measure system, judges assess the items
and award a grade. But a metrically superior item may perform better in this system than an item the judge altogether prefers (colloquially, the whole may be better than the sum of its parts).

Balinski and Laraki reference that, “there are essentially two methods by which social choices can be made, voting, . . . and the market mechanism,” and they say that “the second uses a measure: price expressed in terms of money.” Their paper goes on to advocate for a particular system of voting that they believe is best. These two methods of making social choices represent two characterizations of the role of judges in figure skating competitions. Judges may be understood as voters who pick a winner according to a judging system, or judges may instead be market instruments.

In my study, I assume that judges act on behalf of the ISU and help to pursue its interests. This makes the choice of who wins part of the market mechanism; the winner is the skater that produces a performance most desirable for viewers. But this study illustrates how there are addition layers of complexity about objectively judging performances independent of the judging system that I do not address.

Zitzewitz (2014) studies nationalistic judging bias and vote trading in figure skating competitions. He references his own study of nationalistic bias in the 6.0 judging system that suggested the existence of nationalistic bias and vote trading. In this paper, he examines whether the reforms the ISU made when implementing the IJS affected the instance of bias and found that national favoritism actually increased. Such research into judging bias highlights the importance of the approach to judges as voters independent of the ISU.

Previously, I described the challenge for the ISU being to select a judging system that is fair to skaters and incentivizes desirable performances. Literature related to figure skating
addresses the challenge of selecting a judging system that is fair to skaters. My research employs approaches from contract theory to address the challenge of the judging system incentivizing desirable performances. Aside from how well judges perform their role, how can the judging system itself be optimized?

6 Theoretical Model

This section establishes the relationships between the factors that determine viewership for the ISU and points for the skaters. It yields results of the desirability effect and the difficulty effect that can inform the direction of modifications to points assignments to make a judging system more optimal.

6.1 Introduction to the Theoretical Model

In the IJS case study of the question of optimizing points systems in judged sports, there is a principal and many agents. The principal is the ISU that cares about the viewership \(V\) of figure skating. The agents are skaters who compete in figure skating competitions. Skaters care about success at competitions, which is determined by points \(\rho\). The viewership the ISU receives is a function of the performances \(x\) of the skaters, and the performances skaters give are functions of their allocations of training effort \(\phi\). The ISU awards skaters points \(\rho\) as a function of their performance \(x\). As stated, performance is a function of training effort \(\phi\). The contract between the principal and the agents is the points system through which the ISU awards points.

When I talk about performance with the vector \(x\), I am referring to every aspect of a
program a skater performs that viewers receive. The principal chooses to identify aspects of this performance to evaluate. For example, the ISU might choose to index one aspect of performance as the axel jump, another aspect of the performance as a spin element, and another aspect of the performance as the success of the choreography at portraying a character. Such indices would allow judges to evaluate the performance of individual aspects of a program as the make up the overall performance. In this model, elements of performance are indexed as $x_s$ to indicate the performance of any of the $S$ elements the ISU might decide is relevant to the sport. But I will refrain from suggesting what these $S$ aspects of performance should be. I consider the case of two elements, $S = 2$. My results are not qualitatively changed if I allow for more than two elements.

Likewise, when I talk about effort with an unindexed $\phi$, I am referring to total training effort allocation. Effort may be indexed as $\phi_s$ to refer to the effort allocated to any of the $S$ aspects of performance the ISU identifies. Effort is the endogenous variable; skaters select training effort. Thus, the model is concerned with the contract that induces skaters to optimize points by selecting effort allocation that yields the performance that also optimizes viewership. The performance that optimizes viewership for the ISU and points for the skater corresponds to an optimal effort allocation. Label this optimal effort allocation $\phi^*$. There are several key assumptions about effort that must be noted.

**Assumption 1** An optimal effort allocation $\phi^*$ exists.

**Assumption 2** $x$ is concave.

Without further restriction, there may be multiple optimal $\phi^*$. Assuming $x$ is concave means there are decreasing, non-negative returns to effort. Non-negative returns to effort
implies that more effort will always improve performance, which is reasonable.

**Assumption 3** \( \phi \) is *constrained*.

Since more training effort will always improve performance, a points maximizing skater with unconstrained training effort will allocate infinitely high training effort to improve performance. This is an absurd result.

**Assumption 4** \( V \) is *concave*.

There are decreasing, non-negative returns of performance to viewership. Better performances will always yield greater viewership.

The key implication of these conditions is a unique \( \phi^* \); there exists one allocation of training effort that yields higher viewership than any other allocation subject to the constraint on effort.

To understand how to go about constraining \( \phi \), consider what factors it might include. There are limited hours in a day that a skater might be able to allocate to training. There are limited financial resources that a skater may be able to invest. There is limited access to the ice rink and limited time coaches may have available to dedicate to a skater. \(^1\)

Of course, a skater may not choose to use all resources that could possibly be dedicated to training. For example, although there are 24 hours in a day, a skater most likely would choose to use some of these hours to sleep or take personal time for other pursuits. Let the \( \phi \) in the model be only the training resources that are available to skating. Therefore, all available training resources are used in the optimal performance, because not using all

\(^1\)In comparable sports, training resources are not identical to figure skating. Let all training resources whatever they may be for any sport be implicit in \( \phi \).
resources would yield an internal solution, which would mean a performance that is Pareto inferior.

I could arbitrarily select any value to represent all training resources. I will use $\phi = \phi_1 + \phi_2 + \ldots + \phi_S = 1$ so that each $\phi_s$ corresponds to the proportion of total training effort allocated to an element.

**Assumption 5** $\phi_i \geq 0$.

That is to say that a skater can never allocate negative training effort to a skill and apply that effort to a skill that earns them more points. If the optimality condition is met at a negative $\phi_s$, this simply implies there is a corner solution.

### 6.2 Full Information Points-Maximization

Consider an initial model of full information points-maximization where there are two aspects of figure skating performance. Skaters select training effort allocation to maximize points. Suppose there is no randomness ($\theta$). This means that training effort allocation perfectly determines performance. The ISU’s problem is:

$$\max_{\phi} V(x(\phi)), \text{ s.t. } \phi \text{ maximizes } \rho(x(\phi)), \text{ and } \phi_1 + \phi_2 = 1.$$  \hspace{1cm} (1)

We can write $\phi_2 = 1 - \phi_1$, and substitute so that $x(\phi_1, \phi_2) = x(\phi_1, 1 - \phi_1)$.

For simplicity, denote $\phi_1$ as $\phi$.

When a performance includes two skills, $x_1$ and $x_2$, the skater’s problem is to select the level of effort $\phi_{skater}^*$ that maximizes points.
\[ \rho(x(\phi)) = \rho(x_1(\phi), x_2(\phi)). \] (2)

When there is an interior solution, points are maximized when the derivative of points with respect to effort is equal to 0.

\[ \frac{d\rho}{d\phi} = \frac{d\rho}{dx_1} \frac{dx_1}{d\phi} + \frac{d\rho}{dx_2} \frac{dx_2}{d\phi} = 0 \] (3)

Note that for \( x_2 \), the coefficient will be negative because performance of the skill is a function of \( 1 - \phi \).

Return now to the ISU’s problem to maximize viewership. The optimal level of effort \( \phi_{ISU}^* \) that will maximize viewership is represented by the optimality condition:

\[ \frac{dV}{d\phi} = \frac{dV}{dx_1} \frac{dx_1}{d\phi} + \frac{dV}{dx_2} \frac{dx_2}{d\phi} = 0 \] (4)

Therefore, we have expressions for \( \phi_{skater}^* \) and \( \phi_{ISU}^* \). The problem remaining is to assign a points system whereby \( \phi_{skater}^* = \phi_{ISU}^* \); call the jointly optimal level of effort \( \phi^* \).

In an additive points system with two elements, total points as a function of performance is a coefficient times the performance of the first element plus a coefficient times the performance of the second element:

\[ \rho(x) = p_1 x_1 + p_2 x_2 \] (5)

Select \( p_1 \) and \( p_2 \) so that \( \phi = \phi^* \). Under this condition, the points system induces skaters to
maximize points by selecting effort that also optimizes viewership for the ISU. This optimal
effort corresponds to an optimal performance:

\[ x^* = x(\phi^*) \]  \hspace{1cm} (6)

Note that the magnitude of the points is irrelevant; it does not matter whether skaters earn 10 points or 100 points. All that matters is whether the points consistently and accurately reflect the value, where value is determined by the viewers performances attract. Therefore, the points awarded for an element should be a function of the viewers attracted by the performance of that element, and the points awarded to element \( s \) in the optimal performance should be determined by:

\[ \frac{d\rho}{dx_s} = \frac{dV}{dx_s}(x^*) \]  \hspace{1cm} (7)

Finally, we can set:

\[ p_1 = \frac{d\rho}{dx_1}, \quad p_2 = \frac{d\rho}{dx_2} \]  \hspace{1cm} (8)

Thus, the additive points system \( \rho(x) = p_1x_1 + p_2x_2 \) causes the skaters’ points optimization problem and the ISU’s viewership optimization problem to be solved at the same \( \phi^* \).

6.3 Incomplete Information Points-Maximization

Now suppose there is randomness \( \theta \). To conceptualize the role that randomness plays in competitive figure skating, consider the fact that skaters fall. They do not intend to fall,
because falls incur points deductions and skaters are points maximizers. The model in which
performance is perfectly determined by training effort allocation suggests that skaters can
know exactly how they will perform a program. If a skater knew that a particular effort
allocation would result in them falling, they would have selected a different $\phi$. Thus, the
fact that skaters do fall implies this simple model is insufficient. Therefore, attribute the
factors that skaters do not know or control to randomness, $\theta$.

Including randomness means that $x$ is a function of $\phi$ and $\theta$:

$$x = x(\phi, \theta)$$ (9)

The ISU’s problem is to maximize the expected value of viewership:

$$\max_{\rho} \mathbb{E}[V(x(\phi, \theta))], \text{ s.t. } \phi \text{ maximizes } \mathbb{E}[\rho(x(\phi, \theta))], \text{ and } \phi_1 + \phi_2 = 1.$$ (10)

Once again, we can write $\phi_2 = 1 - \phi_1$, and substitute so that $x(\phi_1, \phi_2, \theta) = x(\phi_1, 1 - \phi_1, \theta)$,
and denote $\phi_1$ as $\phi$.

When a performance includes two skills, $x_1$ and $x_2$, the skater’s problem is to select the
level of effort $\phi_{skater}^*$ that maximizes the expected points.

$$\mathbb{E}[\rho(x(\phi, \theta))] = \mathbb{E}[\rho(x_1(\phi, \theta), x_2(\phi, \theta))].$$ (11)

Expected points are maximized when the derivative of points with respect to effort is
equal to 0.

\[
\frac{d}{d\phi} \mathbb{E}[\rho(x_1(\phi, \theta), x_2(\phi, \theta))] = \mathbb{E}\left[\frac{d}{d\phi} \rho(x_1(\phi, \theta), x_2(\phi, \theta))\right] = \mathbb{E}\left[\frac{d\rho}{dx_1} \frac{dx_1}{d\phi} + \frac{d\rho}{dx_2} \frac{dx_2}{d\phi}\right] = 0.
\]

(12)

Likewise, the optimal level of effort \(\phi_{ISU}^*\) that will maximize viewership is represented by the optimality condition:

\[
\frac{d}{d\phi} \mathbb{E}[V(x_1(\phi, \theta), x_2(\phi, \theta))] = \mathbb{E}\left[\frac{d}{d\phi} V(x_1(\phi, \theta), x_2(\phi, \theta))\right] = \mathbb{E}\left[\frac{dV}{dx_1} \frac{dx_1}{d\phi} + \frac{dV}{dx_2} \frac{dx_2}{d\phi}\right] = 0.
\]

(13)

Again, assign a points system whereby \(\phi_{skater}^* = \phi_{ISU}^*\) and call the jointly optimal level of effort \(\phi^*\). As in equation (6), this optimal effort allocation corresponds to an optimal performance. The task is to select \(p_1\) and \(p_2\) (as in equation (5)) so that skaters maximize expected points at the optimal expected performance for the ISU. The points awarded to element \(s\) in the optimal expected performance should be determined by:

\[
\frac{d\rho}{dx_s} = \mathbb{E}\left[\frac{dV}{dx_s}(x^*)\right]
\]

(14)

Then set:

\[
p_1 = \frac{d\rho}{dx_1}, \quad p_2 = \frac{d\rho}{dx_2}
\]

(15)

The resulting additive points system assigns points to performances so skaters who maximize their expected points allocate training effort such that the ISU’s expected viewership is optimized.
6.4 Competitive Agents

So far, theoretical analysis has assessed agents as points maximizers. However, I discussed earlier that skater utility is a function of points in relation to other skaters’ points. There is a strategic interaction problem. Skaters may not need to necessarily maximize points, but only ensure that their points are above all other skaters. This section considers order statistics theory.

Assume that skaters care what place they come; placing first yields a higher utility than placing second, which yields a higher utility than placing third, and so on. Let $\nu_m$ be the utility a skater receives from coming in $m^{th}$ place when there are $n$ opponents and $m \leq n$. The probability that a skater finishes in $m^{th}$ place is $P_m$.

Let $\vec{\phi}$ be the effort allocation vector $(\phi_1, \ldots, \phi_S)$ where each component 1 through $S$ represents the effort allocated to the corresponding skill performed. Let $J(\phi)$ be the probability $\vec{\phi}$ is sufficient to beat one opponent. Thus, $J(\phi)^n$ is the probability that $\vec{\phi}$ is sufficient to beat $n$ opponents. To come in second place (lower utility than first, but higher utility than third), the skater must beat $n - 1$ opponents and lose to one opponent. There are $n$-choose-$1$ ways that this can happen; there are $n$ opponents to potentially lose to. So the probability the skater comes in second is $J(\vec{\phi})^{n-1}(1 - J(\vec{\phi}))n$.

In general, the probability of the skater coming in $m^{th}$ place is:

$$P_m = \Pr(m^{th} \text{ place}) = \binom{n}{m-1} J(\vec{\phi})^{n-(m-1)} (1 - J(\vec{\phi}))^{m-1}.$$  \hspace{1cm} (16)
Total skater utility $\nu$ is:

$$\nu = \sum_{m=1}^{n} \nu_m P_m \quad (17)$$

Substitute in equation (16). Then, the utility maximization problem can be expressed as:

$$\max_{\phi} \sum_{m=1}^{n} \nu_m \left( \frac{n}{m-1} \right) J\left( \phi \right)^{n-(m-1)} \left( 1 - J\left( \phi \right) \right)^{m-1}, \text{ s.t. } \phi \geq 0 \text{ and } \sum_{m=1}^{n} \phi = 1. \quad (18)$$

Utility is maximized when the derivative with respect to effort is equal to zero. This is also the point where the marginal benefit of allocating effort to skills is equal to the marginal cost of shifting effort allocation.

Differentiating equation (18) with respect to effort yields:

$$\frac{dJ\left( \phi \right)}{d\phi} = \frac{\lambda}{\Gamma\left( \phi \right)} \quad (19)$$

With:

$$\Gamma\left( \phi \right) = \sum_{m=1}^{n} \nu_m \left( \frac{n}{m-1} \right) \left[ (n - (m-1)) (1 - J\left( \phi \right)) - (m-1) J\left( \phi \right) \right] J\left( \phi \right)^{n-m} (1 - J\left( \phi \right))^{m-2}. $$

Marginal benefits are $\Gamma(\phi)$ times $dJ(\phi)/d\phi$ and marginal costs are $\lambda$.

This result means that skaters select effort allocation such that $dJ(\phi)/d(\phi_s)$ is equal to a constant across all $S$ elements. The term $dJ(\phi)/d(\phi_s)$ is the change in the probability that $\vec{\phi}$ is sufficient to beat $n$ opponents for a one unit change in $\phi_s$. Intuitively, it makes sense that this should be set to a constant across all elements. Because if this value was not
constant across all elements, the skater could increase their probability of beating opponents by shifting effort away from elements with lower values of $dJ(\phi)/d(\phi_s)$, and towards elements with higher values.

The ISU can set points. If $\rho_s = 0$, then the skater should put no effort into that element. If $\rho_s$ is very high. The skater should put all effort into that element. The optimum distribution of points should be somewhere in between. Thus, this consideration of order statistics yields the conclusion that the ISU can set points in such a way that competitive skaters who optimize their probability of beating other skaters distribute effort to elements that the ISU desires.

7 Theoretical Results

This section discusses the implications of the full information points-maximization model on how a points system should be selected to solve the principal-agent problem.

7.1 Desirability

A desirable performance is one that is good for viewership. Element $s$ is desirable if the rate of change of viewership for a one unit change in the performance of $s$ is high. That is to say, $\frac{dV}{dx_s}$ is high. Recall from equation (7) that the optimization problem is solved at:

$$\frac{d\rho}{dx_s} = \frac{dV}{dx_s}(x^*)$$

This means that the change in points for a one unit change in the performance of $s$ should equal the change in viewership for the performance of that element at the optimal performance. Thus, this relationship yields the result that when element $x_s$ is more valuable for viewership, it should be awarded more points.
This is the desirability effect. All else equal, a more desirable element in terms of its impact of viewership should be given more points. A less desirable element should be given fewer points. In this relationship, it is possible for \( \frac{dV}{dx_s}(x^*) \) to be negative. This would be equivalent to performing something that has a negative impact on viewership, such as mistakenly falling. The desirability effect implies that more undesirable elements should be given greater deductions, and less undesirable elements should be given smaller deductions.

### 7.2 Difficulty

A difficult element is an element that requires more effort to reach the same quality of performance as a less difficult element. This is the same as saying there is a smaller change in performance per unit of effort. When element \( s \) is more difficult, \( \frac{dx_s}{d\phi} \) is smaller. Recall from equation (6) that \( x^* = x(\phi^*) \). A smaller \( \frac{dx_s}{d\phi} \) must mean a smaller \( x^* \). Essentially, all else equal a more difficult skill results in a worse optimal performance because effort is constrained and the difficult skill has lower returns of performance to effort. Equation (7) says that a worse performance should be given fewer points, so fewer points should be awarded to element \( x_s \).

This is the difficulty effect. All else equal, a more difficult element in terms of the returns of performance to effort should be given fewer points because the overall performance will have a smaller benefit on viewership. An easier element should be given more points.

To understand how this might be true, consider this example. Perhaps viewers like to see quadruple jumps because they are very impressive. It may be that viewers are indifferent to whether skaters perform a quad toe-loop or a quad lutz because they find them equally
impressive. A lutz is considered to be a more difficult element than a toe-loop because the entry requires reversing the direction of rotation, whereas the entry for the toe-loop helps to initiate rotation in the right direction. This means it is more difficult to land a quad lutz than a quad toe-loop. If viewers are really indifferent to what sort of quad a skater performs, the ISU should award more points to the toe-loop to incentivize skaters to allocate training effort to the toe-loop. Since it is easier, their training will have greater returns to performance and skaters will perform better (more skaters will land it than the lutz and the quality will be better). This outcome of better performances improves viewership. The ISU currently awards more points to the lutz than the toe-loop. Of course, more enthusiastic members of the audience may know the lutz is harder and desire to see it. There must be a balance.

7.3 Applications

In practical applications, more difficult elements are often better for viewership because they are more impressive. This would mean that the desirability effect and the difficulty effect act in opposite directions. Tabb (2018) explains some of the concerns people have about quadruple jumps in skating. The high difficulty of the jumps means they can have a negative effect on performance. Simultaneously, they are very impressive and exciting to watch. Ford (2018) describes Nathan Chen’s performance at the 2018 National Championships where he landed five quadruple jumps, more than anyone else in the world had landed in a competition. The article expresses excitement to see what he would produce at the
THEORETICAL RESULTS

upcoming 2018 Olympics (a record six quads)\(^2\). Viewers are excited to quadruple jumps; they are desirable.

In the case of quadruple jumps in general, the desirability effect says to give them more points. But the difficulty effect says to give them fewer points. The challenge for the ISU is determining which effect should be given greater consideration for any particular element.

The difficulty effect to give more difficult elements fewer points seems initially counter-intuitive; ordinarily, it would seem reasonable that if someone were to put a lot of effort into doing something difficult, they should be rewarded for it. However, there are examples of judged performance sports deciding to allocate points in this way. For example, at the 2019 world gymnastics championships, Simone Biles performed a new dismount from the beam: a double-twisting double-backflip. This was a more difficult element than anyone had ever competed before.

Armour (2019) writes in an opinion article how the gymnastics federation penalizes Biles for being too good by not awarding enough points for this difficult skill. However, the gymnastics federation explains, “in assigning values to the new elements, the [women’s technical committee] takes into consideration many different aspects; the risk, the safety of the gymnasts and the technical direction of the discipline. There is added risk in landing of double saltos for beam dismounts (with/without twists), including a potential landing on the neck.” This explanation cites the difficulty effect; the difficulty of the skill means gymnasts won’t perform it as well, and performing it may risk injury, which can prevent gymnasts from continuing to compete.

\(^2\)The 2020 national championships were in Greensboro, North Carolina. In the preceding week, Chen practiced at the rink where I coach. His jumps are very exciting to see in person.
Thus, in applications of this theoretical model, it is necessary to consider both the desirability effect and the difficulty effect. These effects may work in opposite directions and it may be necessary to determine which one should be given greater consideration.

8 Empirical Model

In 2004, the ISU changed from using the 6.0 judging system of ranking to the IJS additive points system. In the theoretical model section, I established that it is possible to select an additive points system such that skaters perform programs that are optimal for the ISU. The empirical section of this paper seeks to answer whether the IJS — the additive points system the ISU selected — is the optimal system by analyzing data from before and after the change in the judging system to determine whether performances are statistically different from each other.

The null hypothesis is that performances after the change in the judging system do not differ from performances before the change. The alternative hypothesis is that performances after the change differ from those before the change.

8.1 Relation to the Theoretical Model

I use the number of errors in a performances as the metric to determine whether performances are different or not (I elaborate on this dependent variable later in this section). Thus, the null and alternative hypotheses relate to whether the number of errors in performances before and after the change are significantly different.

For the purpose of this analysis, I assume that programs performed under the 6.0 system
were optimal. I can make this assumption because the 6.0 system was entirely based on holistic impressions of performances. As discussed in the background section, judges acting in the interests of the ISU could award first place to programs that were optimal for viewership. The fact that skaters seek to be ranked in first place implies that skaters would select optimal programs to perform or else risk being undercut by another skater who did select an optimal program.\(^3\)

Consider how this empirical model relates to the theoretical model I established and the desirability and difficulty effects. Based on the assumption that performances in the 6.0 system were optimal, a statistically significant difference in the number of errors for performances under the IJS would mean performances were suboptimal and the IJS is not the optimal judging system.

If the number of errors under the IJS is statistically significant and positive, this means skaters made more errors. This result would imply an overvaluation of the desirability effect and an undervaluation of the difficulty effect because it would mean the judging system provided a points incentive for skaters to allocate their constrained training resources to elements that had high returns of viewership to performance, but low returns of performance to effort.

An example of this case might be that there is a very difficult element that viewers love, and the ISU awards so many points to it that performing it makes a skater very likely to win. In consequence, many skaters try it but very few succeed, and all the rest make many errors. This problem could be addressed by placing relatively more consideration on the difficulty

\(^3\)There are various reasons why this assumption could be challenged. For example, as numerous other studies have analyzed, judges may fail to act purely in the interests of the ISU and may act in the interests of their national skating federation. Indeed, this possibility motivated the change in the judging system. However, skaters would still have to assume judges act in the interests of the ISU.
effect and reducing the points awarded to elements for difficulty. Assuming errors are bad for viewership, applying greater deductions to errors would also address this problem.

If the number of errors under the IJS is statistically significant and negative, this means skaters made fewer errors. This result would imply an undervaluation of the desirability effect and an overvaluation of the difficulty effect. It would mean the judging system provided a points incentive for skaters to allocate training resources to elements with high returns of performance to effort and low returns of viewership to performance.

An example of this case might be if double jumps and triple jumps are awarded the same number of points. Double jumps are much easier for skaters, but viewers like triple jumps much more. Because elements are awarded equal points and double jumps are easier for skaters, skaters will perform all double jumps. This problem could be addressed by placing relatively more consideration on the desirability effect and increasing the points awarded to elements for desirability.

Alternatively, I could relax the assumption that the 6.0 system is optimal and instead assume that errors are monotonically bad for viewership. In this model, a positive effect of the IJS would mean more errors and a less optimal judging system. A negative effect of IJS would mean fewer errors and imply that the IJS is more optimal than the 6.0 judging system.

I choose to make the former assumption because I think it more accurately reflects competition; I would imagine that in an optimal competition, some skaters would succeed at difficult elements and others would fail and make errors. A competition in which all skaters made no errors might actually be less beneficial for viewership because it is less interesting. However, the alternative interpretation of the empirical approach is also valid and would
require a different application of the desirability and difficulty effects.

8.2 Dependent Variable: Errors

To describe an empirical model, it is necessary to identify metrics to determine whether a program has changed. I select the number of errors in a figure skating performance. I choose this dependent variable because of the relative availability of data on the number of errors and my own understanding of what makes distinguishably different performances.

Ideally, it would be possible to test the effect of the judging system directly on the ISU’s output of interest, whether this is viewership, or profit, or some other output. The first challenge with this approach is the ambiguity of the ISU’s output of interest; the model does not specify and I say viewership only out of personal preference.

The second challenge is with accessing data. If the output of interest is viewership, I do not have access to data about ticket sales, pricing, turnout, or television audience numbers. In contrast the performances themselves that skaters give are accessible and performance is specified in the model. So data about performances are informative about the results of the theoretical model.

There remains the challenge of determining whether potential changes in the number of errors in men’s free skating programs are caused by the change in the judging system. It would be insufficient to compare the performances of current (2020) skaters to performances before the change because this does not control for other factors such as ability, coaching resources, or changes in training patterns that can affect how likely a skater might be to make an error. To account for this challenge, I utilize the argument of a regression discontinuity.
In the empirical model, I compare performances from just before and just after the change in the judging system by the same skaters. Performances just before the change in the judging system are control condition performances, and performances just after the judging change are the treatment condition performances, where the treatment is the IJS. By employing this strategy, I argue that I can control for differences in ability between skaters, changes in training habits, and changes in coaching to identify the effect of the change in the judging system.

8.3 Effect of a Change in the Judging System

I define empirical models as follows:

\[
\text{Model 1: } \chi = \alpha + \beta_1 IJS + \delta_{1,...,n_{skater}} + \epsilon, \quad (20)
\]

where \( \chi \) is the number of errors in a performance, IJS is the categorical variable for the treatment of the new judging system being true, and there is a categorical variable to identify a performance by every skater 1 to \( n \) in the analysis.

In model 1, every independent variable is a categorical variable. This makes it difficult to find meaningful results. This model would be strengthened by the inclusion of continuous variables. One such continuous variable is the age of the skater at the time of the competition. This is a good variable to include because age is a control for the effect that competitive experience might have on errors separate from the identity of the skater. The data spans about one year around the change in the judging system. The effect of this year on the likelihood of making errors might be greater for a younger skater than an older skater.
according to diminishing marginal returns. Therefore, a better model might be:

Model 2: $\chi = \alpha + \beta_1 IJS + \beta_2 age + \delta_1,\ldots,n skater + \epsilon.$ \hspace{1cm} (21)

A final model is useful to test the robustness of model 2. Another factor that could affect the number of errors a skater makes is whether the competition is in the skater’s nation of origin. That is, so skaters have a home advantage? This can be measured by the inclusion of a categorical variable to identify whether the country hosting the competition is the same as the nationality of the skater in the observation.

Model 3: $\chi = \alpha + \beta_1 IJS + \beta_2 age + \beta_3 homecompetition + \delta_1,\ldots,n skater + \epsilon.$ \hspace{1cm} (22)

In these models, failing to reject the null hypothesis would suggest that programs after the change in the judging system are not significantly different from programs before. Rejecting the null hypothesis would imply programs are significantly different, and the additive points system the ISU selected was suboptimal.

8.4 Time Trends

The previous models seek to answer whether performances after the change in the judging system are significantly different from programs before the change. The final model seeks to identify time trends in performances before and after the change. That is to ask, how did performances change in the years before and after the change in the judging system? To
answer this question, I analyze performances data from as early as 1997 and as late as 2014.

The ISU decided to change judging systems following the 2002 Winter Olympics. The IJS was first used at the 2003 Skate America, and all international competitions began to use the IJS from the start of the 2004-2005 season. Ideally, my analysis would consider skating performances in aggregate before and after the change; I would consider international level performances by skaters with multiple appearances at international competitions in a season. Without controlling for the individual skater, I could measure changes in trends across skaters. However, this approach is limited by the availability of data for performances before the change in the judging system. I rely on performances before the change in the judging system being available to view on YouTube. Because videos are more readily available for winning performances and for famous skaters with many career wins, this would introduce a selection bias because observations before the change in the judging system would reflect only the performances that people thought were worth uploading to YouTube, and not less memorable performances that would be necessary for an aggregate view of skating under the 6.0 judging system.

Because of this challenge, I decided to analyze performances from four skaters (Evgeni Plushenko, Li Chengjiang, Brian Joubert, and Stephane Lambiel) who competed for multiple years both before and after the change and for whom videos of most performances under the 6.0 system exist on YouTube. The benefit of this approach is that observations before and after the change are comparable within a skater. A limitation of this approach is that it requires I analyze skaters who were in the same generation of skating; skaters who progressed through their competitive careers at about the same time. Thus, this approach prevents me from measuring the effects of the change on skaters of different ages, or skaters at different
points in their competitive development.

I estimate the following model:

**Model 4:** \( \chi = \alpha + \beta_1 IJS + \beta_4 \text{time to change} + \beta_5 \text{time from change} + \delta_{1,...,n \text{skater}} + \epsilon \) (23)

In this model, time to change and time from change are measures of the slope of the occurrence of errors across time. I define the date of the change in the judging system as 1 July, 2004, and measured the difference in time between this date and the competition date. I divide this variable into two new variables: one that is the difference between the competition date and 1 July, 2004 for 6.0 performances and 0 otherwise, and another that is the difference between the competition date and 1 July, 2004 for IJS performances and 0 otherwise. Thus, their respective coefficients represent the slope of the errors over time. In this model, a significant coefficient for IJS tells us whether performances after the change are significantly different from performances before, and a significant coefficient for the time from the change can tell us whether performances continue to change in the years after the change, and the direction of this change; whether performances get closer or farther from optimal.

**Data**

The IJS is an additive points system; the skater’s final score is the sum of all the points the skater earns for the execution of elements and all deductions from errors. Detailed breakdowns of how scores are added together for every skater are freely available on the ISU
website. From these detailed score sheets it is possible to count the number of errors a skater makes.

The most obvious error is a fall. If a skater falls at any point during the program, it counts as an error for the purpose of this analysis. A skater can make other kinds of errors that do not involve falling. For example, to avoid falling on a jump for which they have not been able to control a one foot landing on a smooth edge, skaters may step out of their landing, they may do turns on one foot to regain balance, or they may put the other foot down accidentally or intentionally to stabilize themselves. When skaters make any of these errors, judges make deductions on the grade of execution for the element. Therefore, I define that the skater has made an error when a majority of judges on the panel award a negative grade of execution for an element.

Elements other than jumps can have negative grades of execution. For example, a stumble in a step sequence or a loss of balance in a spin would receive a deduction and are also defined as errors for this study. ISU communication 2254 summarizes errors that should receive a negative grade of execution (ISU, 2019). Many of the errors are easily identifiable, and I consider them in my analysis. Some errors are not easily identifiable. For example, “does not correspond to the music.”

Finally, sometimes when skaters feel uncomfortable or off balance in their entry to a jump, they abort the jump. This is referred to as popping. When a skater pops a jump, they do fewer than the planned number of revolutions. However, it is not possible to know

---

4I believe my experience qualifies me to identify the clear errors I identified. However, I cannot hope to be qualified to identify such subjective errors without specific judging training. However, where easily identifiable errors are the product of randomness, how well an element corresponds to the music is a product of choreography. I argue the difficulty in identifying such errors does not present a problem because skaters can choose to not perform this error and it should never happen.
what elements a skater planned. Therefore, I define this error as a skater performing a jump of only two revolutions or fewer when it is not in combination with a jump of more than two revolutions. At the elite international level in men’s figure skating, athletes are capable of doing triple jumps (jumps of three revolutions). Programs tend to be comprised of quadruple jumps, triple jumps, and some double axels (which are two and a half revolutions). Men might perform these jumps with a double jump in combination. Under my definition of this type of error, if a skater planned a quadruple jump and performed a triple, or planned a triple axel and performed a double axel, I would not count an error. However, the skater has not performed an element that is uncommon for their level. Thus, this definition of popping a jump might not capture every instance of performing fewer revolutions than intended, but it will not capture false counts of an error.

The 6.0 system of judging is a ranking system. Judges observe the performances and then rank them relative to other skaters. Consequentially, there are not freely available details of what skaters performed and what errors they might have made. Therefore, data collection of performances in the 6.0 judging system are limited to those performances I am able to watch myself on YouTube and count the errors as defined above.\footnote{I have been a competitive figure skater for 14 years and I have been a professional figure skating coach for 5 years. I have attended international seminars organized by the ISU and seminars that explained the judging system.}

In the IJS, the ISU identifies errors that should receive a negative grade of execution. Based on score sheets, I can count errors by the number of elements that receive a negative grade of execution and the number of popped jumps. I can identify the same errors in videos of 6.0 performances. So observations before and after the change in the judging system measure the same thing.
The data is limited to performances by those skaters for whom there are video performances available from before the change in the judging system and who also competed the following season under the IJS. The sample includes performances by Evgeni Plushenko, Stephane Lambiel, Jeffrey Buttle, Emanuel Sandhu, Evan Lysacek, Johnny Weir, and Brian Joubert. Observations from the 6.0 system come from the 2004 Four Continents Championships, European Championships, and World Championships. Observations from the IJS come from the 2005 Four Continents Championships, European Championships, and World Championships and the 2004 Grand Prix of Figure Skating. According to the analysis from the theoretical model, I assume skaters are always points maximizers, so there should not be systematic differences in performances from different competitions.

Note that I use different sets of skaters for Models 1-3 and Model 4. Models 1-3 are based on performances just before and just after the change in the judging system. I analyze skaters for whom there is data available just before and just after the change. In Model 4, I analyze time trends for several years before and after the change in the judging system. I only analyze skaters for whom there are multiple years of data before and after the change. These are different sets of skaters.

There is a potential for bias in the data availability. I am only able to analyze the performances of skaters for whom there is video of 6.0 performances at these competitions. There may only be a video available because they performed well (or poorly). Therefore, the 6.0 observations may systematically be from good performances with few errors. Then comparing good performances to the readily available IJS data which may include performances with more errors might give a false significant positive result to the effect of IJS on errors. I argue this is not the case with these observations because six of the seven skaters listed
make up six of the top seven finishers at the 2005 World Championships. Therefore, these skaters represent skaters who succeeded under the IJS. So, it is not the case that skaters for whom there is video data only performed well at the end of the 2004 season under the 6.0 system. The seventh skater included in the analysis, Evgeni Pluskenko, placed fifth in the short program at the 2005 World Championships, but withdrew from the competition with injury before the free skate. He went on to win the 2006 Olympics. Like the other skaters in my sample, he did not only perform well at the end of the 6.0 system.

This potential for bias in data availability caused me to be especially careful in the data I collected and control for this potential by analyzing only the performances of seven of the most successful skaters. Table 1 summarizes average error rates for the skaters analyzed. The fact that the data come from consistently excellent (low-error) skaters means that the data is narrower in variance than a large sample would be. Thus, the test is conservative; the probability of a significant difference is lower than it might be with a larger sample.

Age at the time of the competition is determined by taking the difference in days between the skater’s birthday and the competition day and dividing by 365.25. This allows for precise ages to differentiate between competitions within one year.

## 10 Empirical Results

Table 2 shows the results of three models. In Model 1, IJS is not significant at the $\alpha > 0.1$ level. This means I fail to reject the null hypothesis that performances were not different before and after the change in the judging system. This is unsurprising given that there are merely 36 observations and all of the independent variables are categorical variables.
In Model 2 in Table 2, the inclusion of age as a continuous variable allows the model to explain the number of errors by competitive experience and then control for differences in ability between skaters and the change in the judging system. Age has a negative coefficient that is not significant at the $\alpha < 0.05$ level ($t = -1.38$). Although not significant, the sign of this coefficient suggests that older skaters — skaters with more competitive experience — fall less than younger skaters.

The variable of interest is IJS. IJS is significant at the $\alpha < 0.1$ level ($t = 1.73$). This result means that I can reject the null hypothesis that performances after the change are not different from performances before. The coefficient is positive ($\beta = 3.287$), which indicates the change in the judging system has a positive effect on the number of errors.

The positive coefficient on IJS implies that the IJS overvalues the desirability effect and undervalues the difficulty effect. The IJS could be made more optimal by giving greater relative consideration to the difficulty effect and awarding fewer points to difficult elements to incentivize skaters to perform elements with higher returns of performance to effort.

In Model 3 in Table 2, the control for the competition being in the nation of origin is not significant ($t = 0.92$), so this does not improve the model. Furthermore, in this model IJS is no longer significant. This is unsurprising given that there are so few observations.

These results allow me to say that IJS performances are suboptimal at the $\alpha < 0.1$ level. This is an informative result. However, there is a ten-percent chance of a type I error — a false rejection of a true null hypothesis. This analysis could be improved by more complete data. A greater number of observations may decrease the variance and lead to a more significant result. Additionally, a sample that is more representative of the diverse pool of skaters would mean a more informative test.
In Model 4 in Table 3, none of the variables of interest are significant at the $\alpha < 0.1$ level.\(^6\) This model fails to reject the null hypothesis that performances after the change do not differ from performances before. This result could be because of too few observations, or it could be a result of the data limitations I noted in the Data section that prevent the analysis from measuring aggregate effects across groups/skaters in skating.

Figure 1 depicts prediction curves of the number of errors for each of the skaters before and after the change in the judging system based on Model 4. It also includes a scatter plot of all observations. This figure illustrates how the values of coefficients and their signs suggest that the number of errors was approximately constant under 6.0 and that the change to the IJS resulted in more errors and increasing errors over time. If significant, these results would indicate that the IJS is becoming more suboptimal.

\section{Conclusion}

This paper addresses the question: how can the ISU select a judging system to award points to competitive skaters such that the optimal performance for skaters also maximizes viewership? I use approaches from contract theory to characterize ISU-skater interactions as a principal-agent problem.

In the theoretical model, when skaters desire to succeed at competitions and the ISU desires to maximize an output of interest that I call viewership. I find that an optimal performance corresponding to an optimal allocation of training effort exists. At this optimal effort allocation, the performance skaters give to maximize their chance of winning is the IJS is significant at the $\alpha < 0.101$ level. The $p$ value of the IJS variable is very close to statistically significant, but definitively not.
same performance that optimizers viewership for the ISU.

The theoretical model also provides interesting results about how points should be allocated to elements in the optimal performance. The desirability effect says that all else equal, more desirable skills should be given more points. More undesirable skills should be given greater deductions. The difficulty effect says that all else equal more difficult skills should be given fewer points. In practice, these effects often work in opposite directions, so it remains necessary to determine how much consideration to give to each of the effects when a sporting federation assigns points.

Having established a theoretical understanding of points based judging systems, I conducted an empirical analysis of figure skating performances before and after the 2004 change in the judging system to determine whether the IJS is an optimal system. The results of the empirical section suggest that the introduction of the IJS significantly changed performances in comparison to performances in the 6.0 judging system.

The coefficient on the IJS variable is positive, which means the introduction of the IJS increased the number of errors in performances. This implies that the IJS overvalues the desirability effect and undervalues the difficulty effect; it incentivizes skaters to perform elements that viewers like, but on which they make more errors — they don’t perform them as well.

The implication of the theoretical model on this empirical results is that the judging system could be improved by giving greater relative consideration to the difficulty effect and decreasing the number of points awarded for difficult elements. Such a change would incentivize skaters to give performances that are more optimal for viewership.

Even with these results, limitations on the data prevent more meaningful analysis and
conclusions. In particular, I limited my data collection to only those skaters for whom performance data before and after the judging system change were available. This allowed me to avoid systematic bias in data availability and control for other factors that might affect the number of errors. But it limited my empirical analysis to a fixed group of skaters probably unrepresentative of skaters in aggregate. It would be reasonable that these skaters who were established and successful in skating before the change in the judging system would not change their behavior much at all because of the marginal costs of changing effort allocation and ability. It may be much more likely that the next generation of skaters – who learn their fundamental skills in the IJS – performs differently as a result of the change in the judging system. But there is much less data for performances in aggregate under the 6.0 system than there is under the IJS.

A more complete empirical analysis would include observations from more skaters before and after the change in the judging system. It would include performances in aggregate from before the change not limited to those videos available on YouTube. With these changes, the greater sample size and the more representative population of skaters would be more likely to yield informative and significant results.

Another limitation of this research is that it does not distinguish between skaters. The theoretical section assumes that skaters are homogeneous; they have identical skills, talent, and ability. To a certain extent, skaters are similar by virtue of sorting themselves into figure skating competitions. Skaters are utility maximizers and they gain utility from placing well at competitions. If a person has very little skill or talent in skating, they might expect very little utility from competing and sort themselves out of competition.

But there are also many ways that skaters are different from one another. Indeed, the
question of viewership and effort allocation becomes complicated when you consider that some skaters may have different talents — they might be able to perform an element with less effort than another skater — and viewers might have different preferences for what elements they like to see. Optimal performances for the ISU may depend on the variable preferences of viewers, the variable talents of skaters, and consequently which skaters viewers most want to see win. This paper does not account for such variability.

Nevertheless, this paper obtains theoretical results that are informative about what should motivate a judging system when skaters and the International Skating Union are rational. These results can be generalized to all judged performance sports and can be used to understand how sporting federations might make decisions about points allocations. It additionally finds evidence that the current judging system, the IJS, is suboptimal and that it may be improved by applying the results of the theoretical model.
# Tables and Figures

Table 1: Summary Statistics of the Number of Errors by Judging System and by Skater

<table>
<thead>
<tr>
<th>Judging System Summary</th>
<th>Mean Errors</th>
<th>Standard Deviation</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0</td>
<td>1.8181</td>
<td>1.8340</td>
<td>11</td>
</tr>
<tr>
<td>IJS</td>
<td>2.52</td>
<td>1.7588</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Skater Summary</th>
<th>Mean Errors</th>
<th>Standard Deviation</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plushenko</td>
<td>1.6</td>
<td>2.0736</td>
<td>5</td>
</tr>
<tr>
<td>Buttle</td>
<td>3.2</td>
<td>2.2804</td>
<td>5</td>
</tr>
<tr>
<td>Sandhu</td>
<td>4</td>
<td>1.2649</td>
<td>6</td>
</tr>
<tr>
<td>Lysacek</td>
<td>1.75</td>
<td>1.5</td>
<td>4</td>
</tr>
<tr>
<td>Weir</td>
<td>0.8</td>
<td>0.8367</td>
<td>5</td>
</tr>
<tr>
<td>Joubert</td>
<td>2.1429</td>
<td>1.6762</td>
<td>7</td>
</tr>
<tr>
<td>Lambiel</td>
<td>2.25</td>
<td>0.9574</td>
<td>4</td>
</tr>
</tbody>
</table>

| Total                  | 2.2         | 1.8439             | 36           |

Data sourced from ISU detailed judges’ scores and YouTube videos of performances
Table 2: The Effect of Judging System on the Number of Errors

<table>
<thead>
<tr>
<th></th>
<th>Model 1 (Errors)</th>
<th>Model 2 (Errors)</th>
<th>Model 3 (Errors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IJS</td>
<td>0.784</td>
<td>3.287*</td>
<td>2.375</td>
</tr>
<tr>
<td></td>
<td>(0.587)</td>
<td>(1.903)</td>
<td>(2.153)</td>
</tr>
<tr>
<td>Lambiel</td>
<td>0.728</td>
<td>-6.047</td>
<td>-3.431</td>
</tr>
<tr>
<td></td>
<td>(1.067)</td>
<td>(5.020)</td>
<td>(5.788)</td>
</tr>
<tr>
<td>Buttle</td>
<td>1.443</td>
<td>1.830*</td>
<td>1.542</td>
</tr>
<tr>
<td></td>
<td>(1.012)</td>
<td>(1.034)</td>
<td>(1.084)</td>
</tr>
<tr>
<td>Sandhu</td>
<td>2.348**</td>
<td>8.003*</td>
<td>5.844</td>
</tr>
<tr>
<td></td>
<td>(0.963)</td>
<td>(4.205)</td>
<td>(4.831)</td>
</tr>
<tr>
<td>Lysacek</td>
<td>0.0323</td>
<td>-7.431</td>
<td>-4.748</td>
</tr>
<tr>
<td></td>
<td>(1.069)</td>
<td>(5.509)</td>
<td>(6.252)</td>
</tr>
<tr>
<td>Weir</td>
<td>-0.957</td>
<td>-5.821</td>
<td>-3.901</td>
</tr>
<tr>
<td></td>
<td>(1.012)</td>
<td>(3.662)</td>
<td>(4.228)</td>
</tr>
<tr>
<td>Joubert</td>
<td>0.453</td>
<td>-5.021</td>
<td>-2.984</td>
</tr>
<tr>
<td></td>
<td>(0.933)</td>
<td>(4.071)</td>
<td>(4.648)</td>
</tr>
<tr>
<td>Age</td>
<td>-2.910</td>
<td>-1.850</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.109)</td>
<td></td>
<td>(2.411)</td>
</tr>
<tr>
<td>Competition in Nation of Origin</td>
<td>0.737</td>
<td></td>
<td>(0.804)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.129</td>
<td>63.11</td>
<td>40.38</td>
</tr>
<tr>
<td></td>
<td>(0.793)</td>
<td>(44.92)</td>
<td>(51.42)</td>
</tr>
<tr>
<td>Observations</td>
<td>36</td>
<td>36</td>
<td>36</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
Data sourced from ISU detailed judges’ scores and YouTube videos of performances
* p < 0.1, ** p < 0.05, *** p < 0.01
Table 3: Time Effects of the Change in Judging System on the Number of Errors

<table>
<thead>
<tr>
<th>Model 4 (Errors)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Slope (6.0)</td>
<td>-0.00210</td>
</tr>
<tr>
<td></td>
<td>(0.102)</td>
</tr>
<tr>
<td>Time Slope (IJS)</td>
<td>0.0517</td>
</tr>
<tr>
<td></td>
<td>(0.0568)</td>
</tr>
<tr>
<td>IJS</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>(0.403)</td>
</tr>
<tr>
<td>Lambiel</td>
<td>1.179***</td>
</tr>
<tr>
<td></td>
<td>(0.363)</td>
</tr>
<tr>
<td>Li</td>
<td>1.642***</td>
</tr>
<tr>
<td></td>
<td>(0.337)</td>
</tr>
<tr>
<td>Joubert</td>
<td>0.990***</td>
</tr>
<tr>
<td></td>
<td>(0.332)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.810*</td>
</tr>
<tr>
<td></td>
<td>(0.413)</td>
</tr>
<tr>
<td>Observations</td>
<td>147</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

Data sourced from ISU detailed judges’ scores and YouTube videos of performances

* p < 0.1, ** p < 0.05, *** p < 0.01
Figure 1: Prediction of the Effect of the Change in the Judging System on the Number of Errors in a Performance. Data sourced from ISU detailed judges’ scores and YouTube videos of performances. Number of observations = 147.
References

Abad-Santos, Alexander. 2014. “Why People Think Adelina Sotnikova’s Figure Skating Gold Medal Was Rigged.” The Atlantic.


ISU. 2019. “Search Results Web results ISU 2254 Levels of Difficulty GOE Guidelines.” Online PDF.


Reid, Ron. 2006. “Figure Skating’s New Judging System.” *Encyclopædia Britannica*.

SDFSC. n.d. “The 6.0 System of Figure Skating Judging.”

