Interim Rank Biases of Subjective Performance Evaluation in Contests

Hitoshi Mitsuhashi
Keio University

Jungwon Min
Kyushu University

Although research on contests has shown that the unique characteristics of contests generate specific behavioral patterns between contestants, we know little about the ways in which such characteristics influence evaluators who subjectively assess contestants' performance. We investigate cognitive errors of performance evaluations in contests and propose that the evaluators' subjective performance assessments can be biased and distorted because of evaluators' expectations that contestants perform poorly if they have limited opportunities for upward mobility and face significant threats of downward mobility (i.e., the interim ranking effects). An analysis of figure skating competition data supports our predictions.

INTRODUCTION

The performance of economic actors is sometimes evaluated in settings in which (1) their performance generates hierarchical ranking, and (2) actors can raise their rank only when the rank of other members drops (i.e., zero-sum competition) (Lazear & Rosen, 1981). This form of competition is called a contest and is frequently used in various settings: races for promotions, sales representatives' contests, distributions of bonuses and monetary rewards, innovation contests, contests for business opportunities (i.e., consulting), award winning competitions, and athletes in sporting events (Day, Gordon, & Fink, 2012). Previous studies have demonstrated the effects of interactions between contestants, suggesting that competitive relations with other contestants shape actors' opportunities of upward mobility and threats of downward mobility in hierarchical rankings, providing incentives (or disincentives) for efforts for higher performance in contests (e.g., Brown, 2011).

While some contests rank contestants only with objective measures (e.g., strokes in golf), other contests additionally use subjective measures such as team efforts and demeanor, where contestants' final rankings are partially dependent upon evaluators' subjective assessments. The literature on organizational psychology has long indicated that such evaluations are not free from cognitive biases and errors (Arvey & Murphy, 1999; Bol, 2011; Roberson, Galvin, & Charles, 2007). Previous economic research on tournaments focuses on contestants' behavioral patterns; however, we know little about the ways in which such patterns influence evaluators who subjectively assess contestants' performance. Although there is the likely possibility that evaluators may exhibit some particular biases in rating contestants' performances, which can reflect the behavioral patterns of contestants uniquely found in contests, this issue has remained unexplored.

In this study, we shift our focus from the behavior of contestants to that of evaluators and judges in tournaments, which should be important because our knowledge about how contests operate will be more compelling as evaluators in some contests play the significant role in deciding final hierarchical rankings. This study proposes the idea of *interim rank biases* of subjective performance evaluation in contests: evaluators' subjective ratings in contests are distorted by contestants' opportunities for later upward mobility and threats of later downward mobility that reflect their interim rankings. Evaluators adjust themselves to contextual factors that shape contestants' motivation and sense of threats, and their subjective performance evaluations may be biased to be consistent with these lay hypotheses that they develop. In particular, we predict that evaluators make expectancy-consistent ratings in the subsequent rounds and give lower subjective performance ratings to contestants with greater distances to rivals with higher interim ranks and to those with smaller distances from rivals with lower interim ranks. Using data from figure skating competitions for 2006–2013, we demonstrate interim rank biases in contests.

THEORY AND HYPOTHESES

Contestants' Behavior

We first review the literatures on contestants' behavior and cognitive biases in performance evaluation and synthesize them to develop our hypotheses about interim rank biases. In some contests, contestants compete in a single round and immediately receive their final ranks, whereas other contests consist of multiple rounds and award contestants both interim and final ordinal ranks. In the latter type of contest, interim ranks are tentative or provisional, resting on some partial outcomes of contestants' performances (e.g., the first half of contests).

In contests with multiple rounds, interim ranking in preceding rounds generates potential opportunities of upward mobility and threats of downward mobility in the later rounds and thus in final ordinal rankings (Feri, Innocenti, & Pin, 2013). Research has demonstrated that contestants react sensitively to such opportunities and threats and exhibit two behavioral patterns in contests. First, a number of studies have investigated the effects of opportunities of upward mobility. Research on adverse superstar effects finds that the presence of a super-competent rival or a superstar in a contest (e.g., Tiger Woods in PGA Tour events) is associated with lower performance of other contestants, who expect lower probabilities of winning, thus reducing their work effort and motivation (Brown, 2011; Tanaka & Ishino, 2012). Casas-Arce and Martínez-Jerez (2009) also find that in sales contests organized by a commodity company, contestants decrease their effort levels following an increase in the lead of rivals leading the contests. Actors have lower levels of motivation with decreases in their perceived probabilities of attaining valuable goals such as upward ranking mobility in contests (Vroom, 1964). Using relative performance in interim rankings, contestants assess their probabilities of winning and upward mobility in final ordinal rankings and decide on the amount of effort that they will make in the later rounds of contests. Contestants tend to increase their effort when they have greater opportunities to pass rivals with higher interim ranks or reduce their effort and underperform if such opportunities are limited.

Second, research also examines the effects of threats of downward mobility. Contestants are under more pressure and are more likely to engage in risk-taking behavior when being more closely trailed by rivals with lower interim rankings (Bothner, Kang, & Stuart, 2007). Contestants' "distaste for moving down" exceeds their "desire to move up" (Bothner et al., 2007: 210). This is because the psychological costs of falling in the rankings are higher than the psychological gains of increasing in the rankings even when the degree of upward and downward mobility is the same (Eisenkopf & Teyssier, 2013; Kahneman & Tversky, 1979). In addition, contestants use rivals close to them in the interim rankings as social reference points and assess their gains and losses in contests relative to the final rankings of such contestants. Being passed in the rankings by closely trailing rivals causes psychological distress (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). Hence, when holding a slight lead in a contest, contestants experience greater pressure to protect their current rankings. Using auto race data, Bothner et al. (2007) find that racers who have a chance of losing their current status to rivals are more likely to take risks and more frequently crash in races. In an experimental study, Pettit, Yong, and Spataro (2010) also find that subjects make more cognitive efforts to resolve mathematics problems that they cannot resolve without computers if facing situations of potential status loss than those of potential status gain.

Potential downward mobility in subsequent rounds that occurs when contestants are narrowly trailed by other rivals generates pressure and causes their underperformance (Cottyn, De Clercq, Pannier, Crombez, & Lenoir, 2006). For example, Genakos and Pagliero (2012) use data from weight lifting competitions and find that contestants tend to fail to lift a given weight when competition is intense or when a successful lift is important for improving their competition ranking. This underperformance results from the effects of choking under pressure (Baumeister, 1984; Beilock, 2007). When individuals' cognitive resources for execution are constrained, performance pressure compromises contestants' working memory, which is critical for online processing of information relevant to effective task execution and disrupts their ability to focus on primary tasks for high performance. The availability of cognitive resources transforms pressure into a positive stimulus that increases their performance (e.g., Kocher, Lenz, & Sutter, 2012); it is therefore a critical boundary condition. Pressure creates distracting situations in which the compromised working memory limits the amount of information available to be processed. In addition, performance pressures raise self-consciousness, which increases the likelihood that contestants will monitor the execution of each of their routinized behaviors. When contestants typically execute large parts of their routines without conscious awareness, this monitoring and attending to the internal performance process paradoxically disrupt the execution of well-learned or proceduralized routines and the automatic quality of behavior. Empirical studies largely support this theory's prediction about the detrimental effects of pressure on performance (e.g., Wells & Skowronski, 2012). For example, Apesteguia and Palacios-Huerta (2010) report the negative effects of pressure on performance by using soccer penalty shootout data.

Schema-based Processing and Evaluators

Subjective performance evaluations are not free from cognitive biases and errors. Evaluators are active processors of information who decide which parts of informational environments and cues about contestants they pay attention to and how they encode the cues for later use (DeNisi, Cafferty, & Meglino, 1984). To do so, evaluators activate schemas representing "subjective 'theories' about how the social world operates" (Markus & Zajonc, 1985: 145) or "a cognitive structure that represents knowledge about a concept or type of stimulus" (Fiske & Taylor, 1991: 98). Schemas are interpretative frameworks, preconceived notions, plots, scripts, or preexisting knowledge structures.

Evaluators' schemas function as cognitive filters through which all cues from environments pass, and shape what kinds of cues are attended to and hence what is observed and encoded. For example, Govaerts et al. (2013) examine how evaluators rate the performance of trainees in medical training programs and find that evaluators, activating their own schemas, attend to trainees' task executions or personal traits, depending upon the length of their own experience. In addition, schemas not only facilitate interpretations of actual stimulus information and inferences by filling in missing information about cues (Fielder, 1982), but also function as "pointers" of attention (Smith, 1988). When evaluators' cognitive resources are constrained or when evaluators are unconsciously exposed to valenced information, they are more likely to direct their attention to, and to recall, schema-relevant and -consistent cues (Macrae, Schloerscheidt, Bodenhausen, & Milne, 2002). Moreover, schemas also influence the recalling and retrieving of related events from memory when evaluating contestants' performance in the case of memory-based, rather than online or immediate ratings. Evaluators are more likely to recall contestants' behavior that is consistent with expectations, which prevents schema-inconsistent events from appearing when evaluating contestants. This is because schemas establish structures on disorganized and unstructured cues by which evaluators can efficiently organize cues and retrieve relevant ones while recalling when making evaluations (DeNisi & Peters, 1996).

The schema-based processing in performance evaluation is related to the halo effect (i.e., evaluators who find one positive characteristic of subordinates tend to extend the positive rating to their other characteristics) in the sense that evaluators extend their images of others based on expectations as a substitute for missing information and develop overall assessments. In addition, it is also related to assimilation effects, which occur when individuals shift their judgments in the direction of valenced information (Bless, Fiedler, & Strack, 2004). Assimilation effects are more likely to occur than contrast effects (i.e., judgments and contextual information are negatively correlated) with the increase in cognitive loads (Levin, 2002).

Interim Rank Biases

A synthesis of contestants' behavior and evaluators' schema-based processing suggests that evaluators may have personal theories about the linkages between situational factors and behavioral patterns and activate the schemas for evaluation particularly when evaluators' cognitive resources are constrained. In other words, under such conditions, evaluators may rely on their lay hypotheses that they develop from their past experiences to reduce the cognitive burdens required for subsequent performance evaluations.

Evaluators' experiences of observing contestants' performances lead them to develop their personal theories about how contestants' distances to and from rivals with respectively higher and lower interim ranks influence their performance in the later rounds of contests. As previous research on contestants' behavior shows, if evaluators recognize that contestants facing mobility opportunities and threats exhibit patterned behavior, evaluators will form their lay hypotheses about relations between interim rankings and contestants' subsequent performances. Such knowledge may lead evaluators to expect the contestant's subsequent poor performance, promoting them to interpret actions in negative ways, attending more to unsuccessful parts of contestants' actions, and more frequently recalling negative actions from memory when making evaluations.

One theoretical basis of this argument is priming effects where ex ante information primes individuals and distorts their information processing in such a way that the subsequent interpretation tends to be consistent with the primed information (Petty & Wegener, 1993; Stanovich & West, 1983). We argue that a type of primed information on which evaluators rely is contestants' provisional distances to and from rivals in contests, which ultimately influence overall rankings. Another theoretical basis is performance cue effects, referring to evaluators' tendencies to give positive ratings if given information about positive past performance (Baltes & Parker, 2000). However, the idea that we are proposing is different in the sense that it rests on evaluators' implicit theories about associations of contestants' distances to and from rivals, not past performance, with the subsequent performance evaluations. Nevertheless, all of these ideas share the possibility that knowledge about interim rank may prime evaluators and cause biased evaluations (Baltes & Parker, 2000).

These arguments lead us to propose the idea of interim rank biases, which can be formally stated in the following two hypotheses.

> *Hypothesis 1: A contestant receives lower subjective ratings for the subsequent performance with an increase in his or her performance gap relative to the rivals with higher interim rankings.*
>
> *Hypothesis 2: A contestant receives lower subjective ratings for the subsequent performance with a decrease in his or her performance gap relative to the rivals with lower interim rankings.*

While it is also a plausible prediction that the behavior of contestants who are subject to ranking anxiety is also affected accordingly, we do not hypothesize it because it has been already demonstrated in previous research.

METHODS

Data and Sample

This study uses data for 60 major male and female senior skating competitions organized by the International Skating Union (ISU) in 2006–2013. The skating competition data have been used in studies such as Balmer, Nevill, and Williams (2001). The major competitions in our sample include the Winter Olympics, World Championships, European Championships, Four Continents Championships, and Grand Prix Series. Each competition has two programs (or rounds) in two days: short and free. This study uses the results of the short programs for measuring interim rankings and subjective ratings for skaters' performance execution in free programs as the dependent variable. The maximum performance time for the short program is two minutes and 50 seconds, whereas that for the men's and ladies' free programs is four minutes and 30 seconds, and four minutes, respectively. In both programs, skaters execute skating elements and moves such as jumps, spins, step sequences, and spirals. In short programs, there are seven required elements (e.g., a double or triple axel jump and a flying spin), whereas skaters can choose the elements for the free programs at their own discretion, but receive no credit and are penalized if executing more than a fixed number of each type of element.

In each competition, skaters compete with each other based on total points, which is the sum of the total element score and the program components score. The total element score represents a sum of scores that skaters are awarded for the execution of each element. Because elements require varying degrees of difficulty and skill, each element has base values set by the ISU. A panel of eight to 12 judges then evaluate how well skaters execute each of the elements and award a grade of execution (GOE) score for each of them, taking integer values from –3 to +3. The scores that skaters are awarded for each element are the sum of the base values and the average GOE scores.

The program component scores consist of five elements: (1) skating skills (i.e., a skater's ability to skate), (2) transitions (i.e., variations and quality of footwork, positions, movements, and holds that link all elements), (3) performance/execution (i.e., a skater's physical, emotional, and intellectual involvement in interpreting the intent of the music and choreography), (4) composition/choreography (i.e., the quality of all types of movements according to the principles of proportion, unity, space, pattern, structure, and phrasing), and (5) interpretation/timing (i.e., personal and creative translation of the music to movement on ice) (ISU, 2014). Each judge awards points on a scale from 0.25 to 10.00 with increments of 0.25. The program component scores are the average of each judge's scores. The element scores are more objective than the program component scores in the sense that judges rate technical aspects of skaters' execution of elements by using clear ISU guidelines and typically watching videos of each element after each skater's performance. The program component scores, on the other hand, include aspects of performance execution such as "physical and emotional involvement," "idea," "concept," "design," "expression of the music," "use of nuances," and "character of music." Judges rate program component scores at the end of each contestant's performance, so schematic processing is likely as ratings entail recalling the contestant's execution and are thus based on impressions. Different judges are selected by the ISU for competitions other than those in the Grand Prix Series, for which host countries' skating federations select judges.

The unit of analysis in this study is skater contest. The number of skaters participating in free programs differs across contests. Some contests such as the European Championships have more than 40 skaters, but poorly performing skaters may behave differently from average-performing skaters (e.g., they care less about the final ranking). In our analysis, we use the data of skaters whose interim rankings are between two and nine in each contest. We have 75 male and 80 female skaters in the sample. The mean scores of the contests in which male and female skaters participate are 9.17 and 8.41, respectively. Although we exclude from the main analysis the data of skaters who lead interim rankings and thus do not trail any rival, we analyze subjective ratings of their performance in a separate analysis. As we use one-season-lagged data in creating some of the independent and control variables, only post-2006 data are used for the regression analyses.

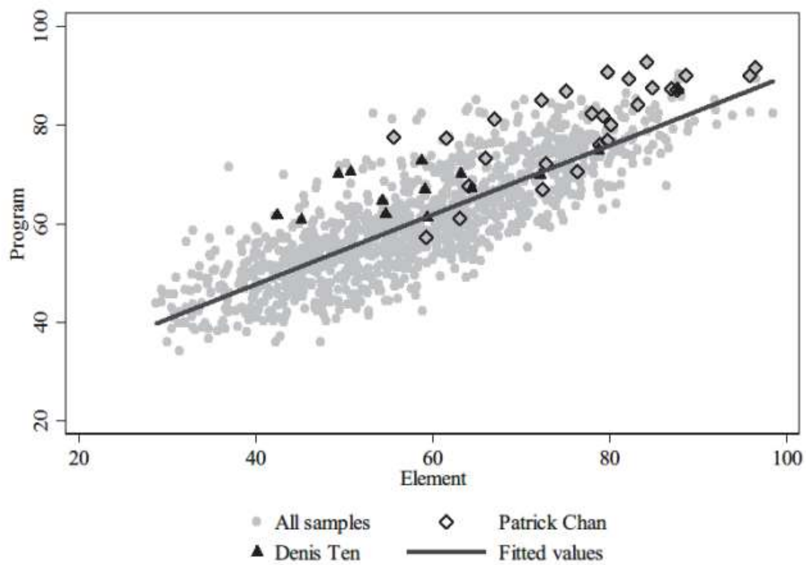
We choose this context for the following reasons. First, both data on judges' subjective ratings and interim rankings that generate opportunities of upward mobility and threats of downward mobility are available. Second, both skaters and judges pay attention to, or at least know, the results of short programs as the interim rankings. Third, the data present a context that fits with our arguments that the evaluation biases occur when evaluators' cognitive resources are constrained. Unlike ratings of element scores for which judges can use videos, their ratings for the program scores are memory based. The narrow intervals between skaters' performances place time pressure on judges, creating situations of constrained cognitive resources. Finally, because only judges with extensive experience are invited to ISU competitions, it is reasonable to presume that judges are familiar with how interim rankings influence skaters' subsequent performances.

Dependent Variables

We use the following two dependent variables for measuring judges' subjective ratings of skaters' performances in free programs. The first dependent variable is skaters' program component scores, which, by definition, rest on judges' overall impressions of the performance and thus should be subjective relative to element scores. Nevertheless, this measure can be controversial if one can claim that, for some reason, skaters actually perform worse (or better) in terms of program component scores with the increase in distance to (or from) rivals with higher (or lower) interim rankings (Roch, McNall, & Caputo, 2011). Therefore, we create the second dependent variable that indicates "deviations" of program scores from the normally expected levels of element scores using the following procedure.

For each gender-season, we run linear regressions to predict program component scores with element scores, compute residuals, and use them as the dependent variable. In other words, the dependent variable captures the variation in program component scores—presumed in this study to be more subjective—that

FIGURE 1
RESIDUALS AND INDIVIDUAL DIFFERENCES



cannot be accounted for by the variation in element scores, which should be more objective. In Figure 1, we pool the male and female skaters' data throughout the observation periods and show the residuals. To check for any individual-level propensity, data for the male gold and silver medalists at the 2013 World Championships (i.e., Patrick Chan and Denis Ten) are marked with triangles and squares, respectively.

Independent Variables

To test hypotheses 1 and 2, we create variables that indicate score differences between a contestant and his/her rivals with a lower or higher interim ranking. Skaters are awarded short program scores immediately after their performance and are hierarchically ranked. *Distance to above* indicates differences in short program scores between a contestant and his/her rival with a higher interim rank, whereas *Distance from below* measures those between a contestant and his/her rival with a lower interim rank. We focus on score differences with rivals with a one-above or one-below interim ranking; in a robustness check, we expand the window by taking summations of the score differences with rivals with a two-above or two-below interim ranking.

Control Variables

We include several control variables concerning characteristics of skaters, judges, and competitions that could influence judges' performance ratings. We include a contestant's *Age* and *Experience* by counting the number of competitions in which a contestant has participated since 2006. We include a variable coded as 1 if a contestant is female. We also include element scores in free programs that may account for variations in the size of residuals and dummy variables that indicate a contestant's interim ranking after the short program. Because judges rate skaters' performances in sequence, skating orders may have some effect (Damish, Mussweiler, & Plessner, 2006). We hence include a variable that indicates skating orders in short programs and one that represents skating orders relative to a contestant's rivals in free programs.

In our context, it is likely that skaters with great performance records are rated more favorably than those with poor performance records (i.e., the halo effect) (Kim & King, 2014). We create the variable *ISU difference with above* by subtracting the ISU points of a rival with a higher interim ranking earned in the preceding season from those that a contestant earned in the preceding season. Skaters are awarded annual "ISU points" based on the accumulation of their final ordinal rankings in each contest. ISU points record skaters' overall performance throughout the preceding season, and are weighted to reflect success in major events such as the Olympics. A larger value of *ISU difference with above* therefore represents the presence of rivals with lower past performance in higher hierarchical rankings, suggesting a greater possibility of contestants' upward mobility. We also create the variable *ISU difference with below* by subtracting ISU points that a contestant earned in the preceding season from those of a rival with lower past performance in the preceding season. Higher values of this variable indicate the presence of rivals with higher past performance in the lower hierarchical rankings, suggesting a greater possibility of contestants' downward mobility. For contestants who did not participate in any contests in the preceding season, we use the lower quartile of the ISU points. The use of the minimum value of the ISU points does not make any significant difference.

In addition, because several studies demonstrate home advantage in sports competitions (Baghurst & Fort, 2008; Balmer, Nevill, & Lane, 2005), we include a dummy variable that we code as 1 if the competition is held in a contestant's home country and 0 otherwise. Because judges tend to favor skaters originating from their own countries (Dawson & Dobson, 2010; Emerson, Seltzer, & Lin, 2009), we also include a variable that counts the number of judges in the panel from a contestant's home country. Because judges who have seen a contestant's performance in the past may rate the contestant favorably (or unfavorably), we include the mean number of times that judges have rated his/her performance since 2006. We also include the number of judges in a panel.

Finally, we include three dummy variables that indicate grades of competition. *WC/OLY* is coded as 1 if the competition is either the World Championships or the Olympics. *EC/FC* is coded as 1 if the competition is either the European Championships or the Four Continents Championships. *GP Final* is coded as 1 if the competition is the Grand Prix Final. We do not include a variable that indicates competitions in the Grand Prix Series. We also include a series of season dummy variables and dummy variables for the focal skater's interim ranks.

Analyses

The unit of analysis in this study is skater-contest. The dependent variable, residual, is continuous. A skater participates in multiple competitions and in multiple season-years. Because the estimation model needs to control for omitted variables that differ between skaters but are time invariant (e.g., skaters' inherent skills and personality) as well as those that differ across competitions but are constant within competitions (e.g., judges' quality), we use random effect regression models, which allow us to incorporate time-invariant variables such as gender and use an unbalanced dataset in which some skaters have only one observation throughout the observation period. A skater may appear several times in the data, so we show robust standard errors of the regression coefficients.

RESULTS

Table 1 shows descriptive statistics and correlations of variables that we use in the main analysis. Some of the coefficients of correlation are high, but the highest and average variance inflation factors are 5.59 and 2.81, respectively. We therefore take no remedial action to correct for multicollinearity.

In Table 2, we present the regression results with control variables to predict the program component scores in models 1 and 2, and residuals in models 3 and 4. Each model includes dummy variables that indicate seasons and the focal skater's interim ranks. Hypotheses 1 and 2 posit that skaters receive lower subjective ratings for their performance if they are further from a rival with a higher interim ranking and are being trailed closely by a rival with a lower interim ranking. In model 1, the coefficient of *Distance to above* is negative and significant, suggesting that skaters' program scores relative to the element scores decrease with an increase in the differences in the short program scores with the rival with a higher interim ranking ($p < .05$). In contrast, we find that the coefficient of *Distance from below* is positive and significant ($p < .01$). The results indicate that skaters' program component scores relative to the element scores decrease with a decrease in the short program score differences with the rival with a lower interim ranking. Hence, the results support hypotheses 1 and 2.

In the analyses above, we exclude the data of skaters who lead the interim ranking because there are no data for *Distance to above*. However, it is informative to examine how their behavior is different from (or similar to) others in competitions because of their importance in contests. There are 142 interim leaders in this dataset. Although running regressions with a small number of observations is an option, we choose to visually inspect the subjective ratings for the leaders' performances in free programs and to focus on *Distance from below*. In Figures 2a and 2b, we box-plot the associations of program scores and residuals, respectively, with *Distance from below* of skaters with the first to fourth interim ranks. We divide the observations using the quartiles of these variables and plot the distributions of residuals for each group. For example, 4Q (i.e., the fourth quartile) and 1Q (i.e., the first quartile) in the figures represent groups of skaters for each interim ranking with scores of *Distances from below* greater than the 75th percentile and lower than the 25th percentile, respectively. It appears from these figures that the leaders' subjective ratings in free programs, whether measured as the sheer scores or residuals, increase when not tailed closely by rivals with lower interim rankings and those with higher past performance. These results are consistent with hypothesis 2 and the results shown in Tables 2 and 3.

TABLE 1
DESCRIPTIVE STATISTICS AND CORRELATIONS

		Mean	S.D.	1	2	3	4	5	6	7	8	9
1	Program scores	60.32	11.48	1								
2	Residuals	–.56	5.49	.44	1							
3	Distance to above	2.29	2.38	.12	–.01	1						
4	Distance from below	2.11	2.11	.17	.07	–.01	1					
5	Age	21.11	3.22	.26	.19	.06	.03	1				
6	Experience	8.21	6.30	.41	.46	.04	.03	.45	1			
7	Female	.51	.50	–.75	.02	–.13	–.13	–.26	–.03	1		
8	Element scores	58.28	13.53	.82	–.05	.12	.15	.12	.19	–.70	1	
9	Interim ranking = 2	.14	.34	.25	.30	.20	.09	.01	.15	–.01	.16	1
10	Interim ranking = 3	.14	.34	.16	.20	.05	.05	–.04	.07	.00	.10	–.16
11	Interim ranking = 4	.13	.34	.07	.12	.02	–.02	.03	.04	.01	.03	–.16
12	Interim ranking = 5	.13	.34	.01	.03	–.04	–.04	–.01	–.02	.01	.01	–.15
13	Interim ranking = 6	.12	.32	–.05	–.04	–.10	–.03	–.02	–.03	.01	–.04	–.15
14	Interim ranking = 7	.12	.32	–.10	–.13	–.05	–.05	–.01	–.05	.00	–.06	–.15
15	Interim ranking = 8	.12	.32	–.18	–.26	–.07	.01	–.01	–.09	–.01	–.09	–.15
16	Short starting order	13.22	11.85	.20	.28	–.11	–.14	–.01	.20	.03	.15	.04
17	Relative starting order in free	1.91	.52	.12	.15	–.03	.00	–.05	.00	.01	.09	.03
18	ISU difference with above	–134.93	1242.95	.09	.20	–.15	–.04	.19	.36	.01	–.01	.01
19	ISU difference with below	–162.92	1152.12	–.14	–.26	–.02	–.12	–.19	–.39	.00	–.03	–.07
20	Home game	.19	.40	–.03	–.07	.03	.03	–.08	–.14	–.02	–.01	–.02
21	Judges from the same country	.79	.41	–.06	–.11	.07	.15	–.02	–.17	–.01	–.03	.03
22	Meeting with same judges	2.20	.81	.37	.49	.00	.00	.29	.78	.05	.17	.16
23	No. of judges on the panel	9.24	.81	–.04	.03	–.03	–.04	.05	–.28	–.12	.02	.00
24	WC/OLY	.12	.33	.26	.26	–.08	–.12	–.03	.24	–.02	.21	–.01
25	EC/FC	.21	.41	.03	.00	–.07	–.09	.02	.01	.00	.05	–.01
26	Grand Prix Final	.05	.22	.15	.16	.05	.16	.05	.15	.00	.13	.08

	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
10	1																
11	–.15	1															
12	–.15	–.15	1														
13	–.14	–.14	–.14	1													
14	–.14	–.14	–.14	–.13	1												
15	–.15	–.14	–.14	–.13	–.13	1											
16	.03	.01	–.01	.02	–.01	–.04	1										
17	.11	.01	.15	.13	–.02	.07	.06	1									
18	–.07	–.01	.00	.06	.03	–.03	.11	–.04	1								
19	–.01	–.01	.05	.03	–.03	.05	–.10	.08	–.50	1							
20	–.02	.03	.02	.00	–.03	.01	–.16	.00	–.06	.08	1						

21	.06	−.01	.01	−.01	−.02	−.04	−.47	.02	−.01	−.04	.18	1					
22	.08	.04	−.03	−.04	−.05	−.08	.44	.00	.30	−.31	−.12	−.30	1				
23	.00	.00	.00	.01	.00	−.01	.18	.01	−.01	.01	−.06	.00	−.25	1			
24	−.01	.00	.00	.01	.00	.00	.61	.01	.02	.00	−.06	−.53	.42	.02	1		
25	−.01	−.01	−.01	.00	.00	.01	.49	.02	.00	.00	−.13	−.12	.15	.12	−.19	1	
26	.08	.08	.08	−.08	−.08	−.08	−.19	.04	.01	−.02	−.06	.11	.12	−.04	−.09	−.12	1

N = 884

TABLE 2
RESULTS OF RANDOM EFFECT REGRESSIONS

Model	1	2	3	4
Dependent variable	Program	Program	Residuals	Residuals
Season dummies	Included	Included	Included	Included
Interim ranking dummies	Included	Included	Included	Included
Age	0.0029	0.006	–0.0085	–0.0051
	[0.10]	[0.09]	[0.09]	[0.09]
Experience	0.1991***	0.2026***	0.2123***	0.2159***
	[0.06]	[0.06]	[0.06]	[0.05]
Female	–13.3912***	–13.3562***	–5.0184***	–4.9912***
	[0.54]	[0.53]	[0.52]	[0.52]
Element scores in free	0.2072***	0.2065***	–0.2777***	–0.2786***
	[0.01]	[0.01]	[0.01]	[0.01]
Short starting order	0.0188	0.0177	0.0218	0.0205
	[0.02]	[0.02]	[0.02]	[0.02]
Relative starting order in free	0.7223***	0.7038***	0.7258***	0.7048***
	[0.20]	[0.20]	[0.20]	[0.20]
ISU difference with above, 1	0.0001	0.0001	0.0001	0.0001
	[0.00]	[0.00]	[0.00]	[0.00]
ISU difference with below, 1	–0.0002*	–0.0002*	–0.0002*	–0.0002*
	[0.00]	[0.00]	[0.00]	[0.00]
Home game	0.5491*	0.5326*	0.4159	0.401
	[0.27]	[0.27]	[0.27]	[0.27]
Judges from the same country	0.7710**	0.7589**	0.8036**	0.7894**
	[0.27]	[0.27]	[0.25]	[0.25]
Meeting with same judges	0.6829**	0.6666**	0.5678*	0.5489*
	[0.24]	[0.24]	[0.25]	[0.25]
No. of judges on the panel	0.2935	0.305	0.2161	0.2282
	[0.21]	[0.20]	[0.20]	[0.19]
WC/OLY	3.9993***	4.1421***	4.1044***	4.2508***
	[0.58]	[0.60]	[0.59]	[0.60]
EC/FC	1.2197**	1.2796**	1.2129**	1.2759**
	[0.39]	[0.41]	[0.39]	[0.41]
Grand Prix Final	2.2111***	2.1131***	2.2666***	2.1612***
	[0.34]	[0.35]	[0.37]	[0.38]
Distance to above		–0.0749*		–0.0894*
		[0.04]		[0.04]
Distance from below		0.1197**		0.1268**
		[0.05]		[0.05]
Constant	40.2440***	40.0037***	6.7528*	6.5430*
	[3.17]	[3.08]	[2.93]	[2.82]
χ^2	2903.7	2972.57	1790.51	1785.54
R^2 within	0.74	0.74	0.58	0.59
R^2 between	0.92	0.92	0.52	0.54
R^2 overall	0.90	0.90	0.60	0.61
No. of observations	884	884	884	884

Robust standard errors are in brackets. * $p < .05$, ** $p < .01$, *** $p < .001$

TABLE 3
RESULTS OF RANDOM EFFECT REGRESSIONS

Model	5	6	7	8
Dependent variable	Element	GOE average	Program	Residuals
Season dummies	Included	Included	Included	Included
Interim ranking dummies	Included	Included	Included	Included
Age	–0.4139*	0.006	–0.0571	–0.0583
	[0.17]	[0.00]	[0.09]	[0.08]
Experience	0.1253	0.0050*	0.2026***	0.2027***
	[0.11]	[0.00]	[0.05]	[0.05]
Female	–19.2318***	0.3930***	–13.1352***	–4.8241***
	[0.88]	[0.03]	[0.56]	[0.57]
Element scores in free		0.0228***	0.2131***	–0.2762***
		[0.00]	[0.01]	[0.02]
Short starting order	0.0113	0.0014	0.0119	0.0155
	[0.06]	[0.00]	[0.02]	[0.02]
Relative starting order in free	0.8279	–0.0069	0.5684***	0.5884***
	[0.64]	[0.01]	[0.14]	[0.15]
ISU difference with above, 1	–0.0002	–1.59E-05	0.0001	0.0001
	[0.00]	[8.57E-06]	[0.00]	[0.00]
ISU difference with below, 1	–0.0001	–1.12E-05	–0.0002	–0.0002*
	[0.00]	[8.31E-06]	[0.00]	[0.00]
ISU difference with above, 2			0.0002	0.0002
			[0.00]	[0.00]
ISU difference with below, 2			–0.0001	–0.0001
			[0.00]	[0.00]
Home game	–0.0386	0.0023	0.7389*	0.5466
	[0.73]	[0.02]	[0.32]	[0.30]
Judges from the same country	2.6735**	0.0227	0.8816**	0.8304**
	[0.84]	[0.02]	[0.30]	[0.29]
Meeting with same judges	–0.2502	0.0089	0.7744**	0.6944**
	[0.82]	[0.02]	[0.26]	[0.26]
No. of judges on the panel	0.5545	–0.0383**	0.4237	0.3132
	[0.68]	[0.01]	[0.22]	[0.21]
WC/OLY	10.2665***	0.0269	4.4966***	4.4039***
	[2.09]	[0.06]	[0.65]	[0.67]
EC/FC	4.4429***	0.0262	1.3412**	1.3012**
	[1.32]	[0.03]	[0.46]	[0.45]
Grand Prix Final	6.4423***	–0.0085	2.4442***	2.4033***
	[1.27]	[0.04]	[0.57]	[0.55]
Distance to above	0.007	–0.003	0.0009	–0.012
	[0.12]	[0.00]	[0.03]	[0.03]
Distance from below	0.2538*	0.0079*	0.0703**	0.0667*
	[0.13]	[0.00]	[0.02]	[0.03]
Constant	58.2859***	–1.2766***	39.2095***	6.2890*
	[8.45]	[0.18]	[2.88]	[2.76]
χ^2	1224.86	1097.59	3073.99	1491.27
R^2 within	0.14	0.49	0.73	0.6
R^2 between	0.80	0.70	0.92	0.55
R^2 overall	0.65	0.57	0.9	0.59
No. of observations	884	884	646	646
No. of skaters	155	155	140	140

Robust standard errors are in brackets. * $p < .05$, ** $p < .01$, *** $p < .001$

As part of a robustness check, in model 5 in Table 3, we use element scores as the dependent variable, which we view to be more objective. The coefficient of *distance to above* is not significant, whereas that of *distance from below* is positive and significant ($p < .05$). The results remain similar when we use means of GOE as the dependent variable in model 6. This indicates a possibility that both skaters and judges react to interim score differences, and that interpretation of the results we report above require some caution as the two effects can be confounded. In models 7 and 8, we use the expanded window and use rivals with two-above and two-below interim rankings in operationalizing the hypothesis testing and control variables. Models 7 and 8 predict the effects on program component scores and residuals, respectively. The coefficients of *Distance to above* are not significant in either model 10 or 11, but those of *Distance from below* are again significant. These results suggest that those regarding downward mobility threats are more consistent and robust.

DISCUSSION

The purpose of this paper is to examine the biases of evaluators in contests and to test the idea of *interim rank biases* in contests in which opportunities of upward mobility and threats of downward mobility that reflect the focal actors' interim rankings distort evaluators' subjective ratings in subsequent rounds. This study proposes that the subjective performance ratings of a contestant decrease with an increase in distance to the rival with a higher interim ranking and with a decrease in distance from the rival with a lower interim ranking. From the analysis of data from figure skating competitions, we find support for these predictions. In particular, our findings for evaluation biases as a result of the threats of downward mobility are stronger and more robust, which is consistent with Baumeister et al.'s (2001) argument that the effects of negative factors such as downward mobility are stronger than those of positive factors such as upward mobility.

To date, our understanding of contests has been limited to how contestants behave in contests. This study shifts the focus from the effects of competitive interactions between contestants on their own behavior to the behavior of evaluators. By drawing on the literature on competitive interactions and cognitive social psychology, and proposing a type of bias found uniquely in the context of contests, our work suggests a new direction in research on the nature of contests as well as evaluators' cognitive biases. Although this study has some commonalities with Eisenkopf & Friehe (2014) in the sense that both studies examine the behavior of noncontestants in contests, our focus on evaluators in contests is new.

As a practical implication, the analysis suggests a risk associated with disclosing interim rankings: contest organizers should not disclose interim rankings if they are interested in avoiding evaluators' cognitive distortions. On the one hand, a number of studies examine the benefits of disclosure in improving the motivation and performance of those participating in contests. Vidal and Nossol (2011) investigate how feedback on relative performance improves the performance of ratees in a German retailer and present two competing predictions: such feedback may serve as a useful benchmark, promote competition, and enhance the motivation of ratees, whereas it can also cause loss of self-esteem and thereby lower motivation. Vidal and Nossol (2011) find an increase in the productivity of ratees after the sample firm starts to disclose performance records for exogenous reasons. In their analysis of a natural experiment in a high school, Azmat and Iriberri (2010) also find that the performance of ratees or students increases when they know whether their performance is below or above average, and how far it is from the average.

FIGURE 2a:
BOX PLOTS OF *PROGRAM SCORES* BY *DISTANCE FROM BELOW* FOR EACH INTERIM RANKING

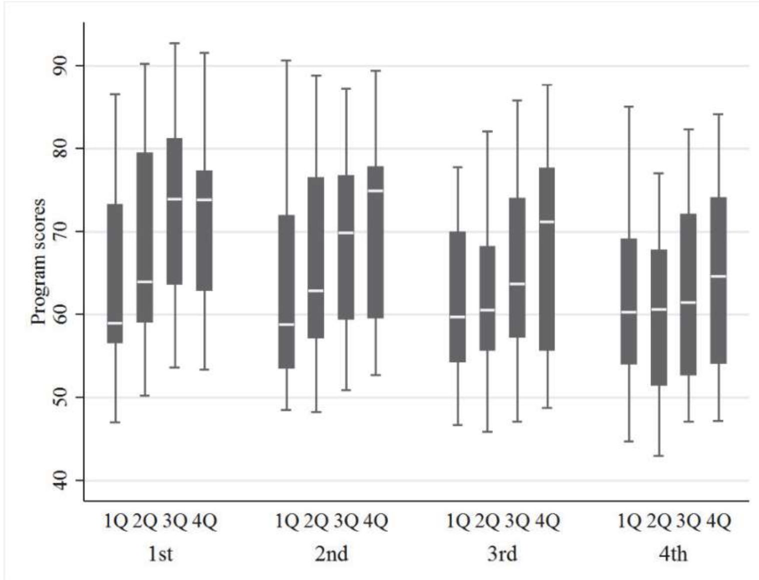
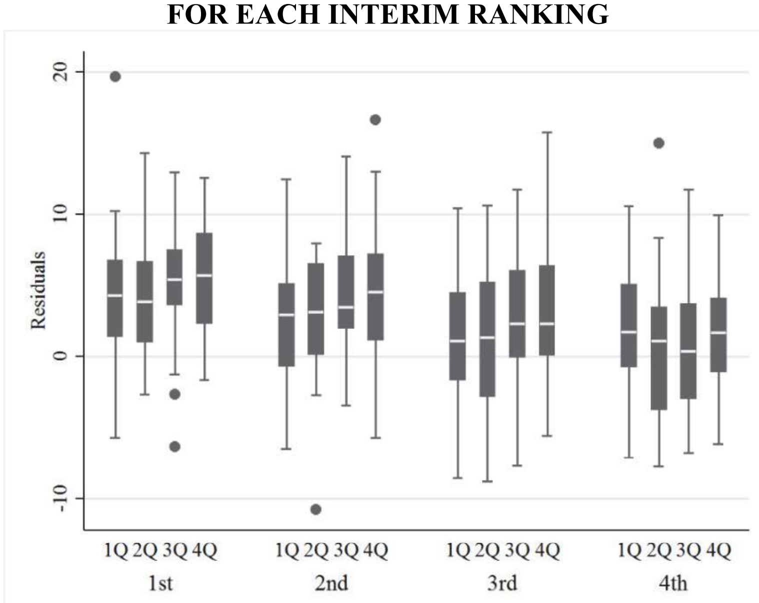


FIGURE 2b:
BOX PLOTS OF *RESIDUALS* BY *DISTANCE FROM BELOW* FOR EACH INTERIM RANKING



Notes: "1st", "2nd", "3rd", and "4th" in Figures 2a and 2b refer to skaters' first, second, third, and fourth interim rankings, respectively. We divide skaters in each rank into four equal groups, 1Q to 4Q, on the basis of *Distance from below*. For example, 4Q (i.e., the fourth quartile) and 1Q (i.e., the first quartile) represent groups of skaters for each interim ranking with scores of *Distances from below* greater than the 75th percentile and lower than the 25th percentile, respectively.

On the other hand, other studies have reported the risk of such disclosure in settings such as professional golf tournaments (Brown, 2011; Tanaka & Ishino, 2012) and weight lifting contests (Genakos & Pagliero, 2012) because contestants' knowledge alters their subsequent behavior and performance. The findings in this study are consistent with those studies in the sense that these studies highlight the risk of disclosing interim rankings. However, unlike previous studies demonstrating the effects of information disclosure on contestants' motivation, the present study focuses on the effect on evaluators. To develop bias-free evaluations, information about interim ranks should be carefully disclosed in contests.

However, further evidence about how information on interim rankings should be disclosed and restricted will be needed before reaching definitive conclusions. Although this study does not explore how the number of rounds in contests influences evaluators' ratings, intuition suggests that the interim ranking effects may weaken in later stages of multi-round contests, where interim rankings may be rather fixed (Vidal & Nossol, 2011). Such investigations may be of importance for contest organizers, who formulate policies indicating at which stages of contests information on interim rankings should be disclosed to minimize the interim ranking effects.

Moreover, it would be interesting to undertake further investigations about the effects of information disclosure and analyze how performance evaluation is biased if interim rankings are disclosed only to evaluators. Even when interim rankings are unknown to contestants, do evaluators who know the interim rankings activate their schema and rate the performance based on potential upward and downward mobility? Does such disclosure increase or decrease the accuracy of performance evaluation? Further empirical efforts in other empirical contexts are needed to address these questions.

A number of limitations of this study also suggest avenues for future research. First, there are several boundary conditions of this research such as constrained cognitive resources available for both skaters and judges (i.e., skaters' consciousness of executing their performance, time pressure for judges, and experience of judges). Future research should examine which of these contextual factors contribute to the findings we report in this study. In addition, we follow the suggestion of Day et al. (2012) to take advantage of athletes' data to advance our knowledge of individual behaviors; however, use of such data clearly constrains the applicability of our results to other contexts, and this is an issue that should be addressed in future research.

Second, future research should presume heterogeneity of evaluators and examine individual-level differences between evaluators. The anonymity of judges' data prevents us from matching judges' scores with their attributes and examining between-judge variations. Thus, our analysis focuses on how contestants are rewarded, rather than on how each judge rewards contestants. This study also limits its scope of evaluators' performance rating, but evaluators may exhibit different types of behavior such as attention and interest. Future research should explore different types of evaluator behavior as a result of competitive interactions between contestants.

Third, the relationships between objective and subjective ratings can be more complex than those presumed in this study. These ratings can be a function of both contestants' "true" performance and performance that judges rate. Using indirect evidence, we believe that our dependent variables capture the more subjective aspects of contestants' performance, but we also obtain a finding that indicates some subjective aspects of the GOE as well. The subjective aspects may not be clearly teased out from the objective ones. Although this study uses field data for the sake of generalizability, it will be an option for future research to overcome these challenges by using experiment data, manipulating "true" or "actual" performance, and computing the difference with subjective assessments. It is also ideal to test our hypotheses in experimental settings, in which, for example, we show subjects as evaluators the same contestants' performance and manipulate the contestants' interim ranking and gaps with higher- or lower-ranked competitors.

CONCLUSIONS

The purpose of this paper is to examine the biases of evaluators in contests and to test the idea of *interim rank biases* in contests in which opportunities of upward mobility and threats of downward mobility that reflect the focal actors' interim rankings distort evaluators' subjective ratings in subsequent rounds. This study proposes that the subjective performance ratings of a contestant decrease with an increase in distance to the rival with a higher interim ranking and with a decrease in distance from the rival with a lower interim ranking. From the analysis of data from figure skating competitions, we find support for these predictions. In particular, our findings for evaluation biases as a result of the threats of downward mobility are stronger and more robust. This study explicates mechanisms by which evaluators' subjective performance ratings are influenced by contestants' interim ranking in contests. This is an important step in helping us to understand how contests work, how evaluators in contests behave, and how evaluators are influenced by competitive interactions between contestants. The findings in this study offer a number of promising research opportunities to integrate various threads in the literature in future work.

ENDNOTES

1. This work was supported by the Japan Society for the Promotion of Science KAKENHI Grant Numbers 25285121 and 16H03658.

REFERENCES

Apesteguia, J., & Palacios-Huerta, I. (2010). Psychological pressure in competitive environments: Evidence from a randomized natural experiment. *American Economic Review, 100*(5), 2548-2564.

Arvey, R. D., & Murphy, K. (1999). Performance evaluation in work settings. *Annual Review of Psychology, 49*, 141-168.

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics, 94*(7-8), 435-452.

Baghurst, T., & Fort, I. (2008). Subjective judging and the home advantage in female collegiate Division I gymnastics. *Women in Sport & Physical Activity Journal, 17*(2), 3-7.

Balmer, N. J., Nevill, A. M., & Lane, A. M. (2005). Do judges enhance home advantage in European championship boxing? *Journal of Sports Sciences, 23*(4), 409-416.

Balmer, N. J., Nevill, A. M., & Williams, A. M. (2001). Home advantage in the Winter Olympics (1908-1998). *Journal of Sports Sciences, 19*(2), 129-139.

Baltes, B., & Parker, C. (2000). Understanding and removing the effects of performance cues on behavioral ratings. *Journal of Business and Psychology, 15*(2), 229-246.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentive on skillful performance. *Journal of Personality and Social Psychology, 46*(3), 610-620.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

Beilock, S. L. (2007). Choking under pressure. In R. F. Baumeister & K. D. Vohs (Eds.), *Encyclopedia of Social Psychology* (pp. 140-141). Los Angeles, CA: Sage.

Bless, H., Fiedler, K., & Strack, F. (2004). *Social Cognition: How Individuals Construct Social Reality*. East Sussex, UK: Psychology Press.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review, 86*(5), 1549-1575.

Bothner, M. S., Kang, J. H., & Stuart, T. E. (2007). Competitive crowding and risk taking in a tournament: Evidence from NASCAR racing. *Administrative Science Quarterly, 52*(2), 208-247.

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy, 119*(5), 982-1013.

Casas-Arce, P., & Martínez-Jerez, F. A. (2009). Relative performance compensation, contests, and dynamic incentives. *Management Science, 55*(8), 1306-1320.

Cottyn, J., De Clercq, D., Pannier, J. L., Crombez, G., & Lenoir, M. (2006). The measurement of competitive anxiety during balance beam performance in gymnasts. *Journal of Sports Science, 24*(2), 157-164.

Damish, L., Mussweiler, T., & Plessner, H. (2006). Olympic nedals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology. Applied, 12*(3), 166-178.

Dawson, P., & Dobson, S. (2010). The influence of social pressure and nationality on individual decisions: Evidence from the behaviour of referees. *Journal of Economic Psychology, 31*(2), 181-191.

Day, D. V., Gordon, S., & Fink, C. (2012). The sporting life: Exploring organizations through the lens of sport. *Academy of Management Annals, 6*(1), 397-433.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*(3), 360-396.

DeNisi, A. S., & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*(6), 717-737.

Eisenkopf, G., & Friehe, T. (2014). Stop watching and start listening! The impact of coaching and peer observation in tournaments. *Journal of Economic Psychology, 45*, 56-70.

Eisenkopf, G., & Teyssier, S. (2013). Envy and loss aversion in tournaments. *Journal of Economic Psychology, 34*, 240-255.

Emerson, J. W., Seltzer, M., & Lin, D. (2009). Assessing judging bias: An example from the 2000 Olympic Games. *The American Statistician, 63*(2), 124-131.

Feri, F., Innocenti, A., & Pin, P. (2013). Is there psychological pressure in competitive environments? *Journal of Economic Psychology, 39*, 249-256.

Fielder, K. (1982). Causal schemata: Review and criticism of research on a popular construct. *Journal of Personality and Social Psychology, 42*(6), 1001-1013.

Fiske, S. T., & Taylor, S. E. (1991). *Social Cognition* (2nd ed.). New York: McGraw-Hill.

Genakos, C., & Pagliero, M. (2012). Interim rank, risk taking, and performance in dynamic tournaments. *Journal of Political Economy, 120*(4), 782-813.

Govaerts, M. J. B., Van de Wiel, M. W. J., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2013). Workplace-based assessment: raters' performance theories and constructs. *Advances in Health Sciences Education, 18*(3), 375-396.

ISU. (2014). International Skating Union Judging System. Retrieved from http://www.isu.org/en/single-and-pair-skating-and-ice-dance/isu-judging-system/introduction

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292.

Kim, J. W., & King, B. G. (2014). Seeing stars: Matthew effects and status bias in major league baseball umpiring. *Management Science, 60*(11), 2619-2644.

Kocher, M. G., Lenz, M. V., & Sutter, M. (2012). Psychological pressure in competitive environments: New evidence from randomized natural experiments. *Management Science, 58*(8), 1585-1591.

Lazear, E. P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy, 89*(5), 841-864.

Levin, A. (2002). Contrast and assimilation processes in consumers' evaluations of dual brands. *Journal of Business and Psychology, 17*(1), 145-154.

Macrae, C. N., Schloerscheidt, A. M., Bodenhausen, G. V., & Milne, A. B. (2002). Creating memory illusions: Expectancy-based processing and the generation of false memories. *Memory, 10*(1), 63-80.
Markus, H., & Zajonc, R. B. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology, 3rd Edition* (pp. 137-229). New York: Random House.
Pettit, N. C., Yong, K., & Spataro, S. E. (2010). Holding your place: Reactions to the prospect of status gains and losses. *Journal of Experimental Social Psychology, 46*(2), 396–401.
Petty, R. E., & Wegener, D. T. (1993). Flexible correction processes in social judgment: Correcting for context induced contrast. *Journal of Experimental Social Psychology, 29*(2), 137-165.
Roberson, L., Galvin, B. M., & Charles, A. C. (2007). When group identities matter. *Academy of Management Annals, 1*(1), 617-650.
Roch, S., McNall, L., & Caputo, P. (2011). Self-judgments of accuracy as indicators of performance evaluation quality: Should we believe them? *Journal of Business and Psychology, 26*(1), 41-55.
Smith, E. R. (1988). Mental representation and memory. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of Social Psychology, 4th Edition* (pp. 391-445). New York: McGraw-Hill.
Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology, 112*(1), 1-36.
Tanaka, R., & Ishino, K. (2012). Testing the incentive effects in tournaments with a superstar. *Journal of the Japanese and International Economies, 26*(3), 393-404.
Vidal, J. B. i., & Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science, 57*(10), 1721-1736.
Vroom, V. H. (1964). *Work and Motivation*. New York: McGraw Hill.
Wells, B. M., & Skowronski, J. J. (2012). Evidence of choking under pressure on the PGA Tour. *Basic and Applied Social Psychology, 34*(2), 175-182.