THE UNIVERSITY OF CHICAGO

The Behavior of Judges under the Gambler's Fallacy

By

Shanshan Liu

June 2021

A paper submitted in partial fulfillment of the requirements for the
Master of Arts degree in the
Master of Arts Program in the Social Sciences

Faculty Advisor: Devin G. Pope
Preceptor: Min Sok Lee

Abstract

Previous studies show that when making sequential decisions, current judgments are affected by proceeding ones and people may exhibit negative autocorrelation. I investigated the behavior of judges when making sequential performance evaluations in competitions to determine if there is a negative autocorrelation that is independent of the performers' merits. I checked its compatibility with the interpretation of the gambler's fallacy, where individuals underestimate the occurrence probability of the sequential streaks. I examined two empirical examples: synchronized swimming competition, and hip hop dance championship, leveraging data from the 2019 FINA World Aquatics Championships, and the 2019 Hip Hop Dance Championships, respectively. The findings suggest that generally, negative autocorrelation is significant and is consistent with the gambler's fallacy, even after procedures such as dropping extreme scores and taking average scores. In these cases, the spontaneous adjustments backfire the goal of fairness. Thus, further improvements such as adjusting the scoring procedures or correcting judges' erroneous beliefs are needed to mitigate the effect.

1 Introduction

Previous studies provided empirical evidence that decision-making is not always rational and deliberate but relies on fast and costless heuristic principles (Tversky and Kahneman, 1974). These heuristic principles include the gambler's fallacy, where individuals wrongly believe that if a particular event occurs more frequently than usual during the past, it is less likely to happen in the future (and vice versa). When judges face sequential evaluations, research showed that current judgments are affected by proceeding ones. People may erroneously negatively auto correlate their actions regardless of the merits of the cases due to heuristics like the gambler's fallacy (Chen, Moskowitz, and Shue, 2016). However, prior studies investigating the sequential judgment biases of judges in competitions all neglected gamblers' fallacy as a potential explanation. Inspired by these findings,

I aim to extend the concept of the gambler's fallacy to competitions, focusing on judges' behavior. I use examples of the synchronized swimming competitions and hip hop dance championships, where the performances are in a randomized order, to test the negative autocorrelation. My research question is:

Do judges experience negative autocorrelation when they sequentially evaluate performances that are compatible with the gambler's fallacy (i.e., would they be less likely to offer a high score following sequential high scores that they experience, independent of the merits of the competitors' performance, and vice versa)?

The results show that there exists a negative autocorrelation that is independent of the performers' merits, which is compatible with the interpretation of the gambler's fallacy, where individuals underestimate the occurrence probability of the sequential streaks. In addition, I find autocorrelations in all three levels (individual, panel, overall) in both examples. The finding suggests that such sequential judgment biases are systematic as they persist even in some cases after procedures such as dropping extreme scores and taking average scores. Since all qualified judges are experienced, this kind of judgment anomalies would hence not attenuate with market experience. Thus, my project has implications for the policy debate on fairness of the competitions. To mitigate the biases, I suggest correcting the erroneous belief before the evaluations or changing the rules, which might be helpful.

This paper adds to the existing literature. Previous research investigating sequential judgment bias has examined examples in both sport competitions such as diving (Kramer, 2017), figure skating (Bruin, 2006), gymnastics (Damisch, Mussweiler, and Plessner, 2006), and non-sport events like idol series (Page and Page, 2010), musical competition (Flôres and Ginsburgh, 1996), song contest (Bruin, 2005), and wine tasting (Mantonakis et al., 2009). Most of these studies have focused on the serial position effect (Bruin, 2005, 2006; Flôres and Ginsburgh, 1996; Mantonakis et al., 2009; Page and Page, 2010), where the judges favor either the first or the last performer. Using the data from gymnastics and synchronized diving competitions, respectively, Damisch, Mussweiler, and Plessner (2006) and Kramer (2017) proved that judges make negative auto-adjustments when

evaluating the performances. But both studies explained such behavior by the sequential contrast effects (SCE), where judges evaluate the performances by comparisons. I intend to complement their research using my examples, as they are similar but concern the compatibility of another explanation of the gambler's fallacy. With multiple judgments consecutively made by the judges, whether they would award scores to the following show to compensate the previous can be found. I discuss further about the difference between SCE and gambler's fallacy in the discussion section. Even if SCE cannot fully distinguish from the gambler's fallacy using my examples, I push the boundaries further as I investigate more subsequent decisions compared to their studies. The orders in my examples are random, and thus, my study is free from the scenarios in their gymnastics and diving examples when judges infer the quality of performances from the starting orders. On the other hand, the above studies have provided convincing reasons for cases where the effect of the gambler's fallacy is insignificant. If the judges tend to seek similarities in performances or are dominated by serial position effects, they will show positive autocorrelations.

As for the gambler's fallacy, Tversky and Kahneman (1974) introduced that, when people make decisions, especially the subjective ones, they rely on heuristic principles which reduce the complex tasks of predicting values to simpler judgmental operations. These simpler judgmental operations could sometimes lead to severe and systematic errors. Individuals show the misconception of chance (i.e., the gambler's fallacy), where people wrongly believe even small samples are highly representative of the populations and underestimate the likelihood of the event that occurs more frequently in the past. Followed by these findings, Chen, Moskowitz, and Shue (2016) worked on three empirical pieces of evidence to show that people underestimate the probability of sequential streaks and negatively auto correlate their actions under such heuristics when making high-stakes decisions based on their discretion. Paralleled to their study, I test whether such a mechanism also affects the competition results, with the order of evaluations guaranteed to be random. I extend further to see if its influence diminished with procedures such as dropping extreme scores and taking average scores. If not, then multiple judges in the competition are equivalent to one single judge from the result perspective, which is the actual case given my findings of no mitigation. Thus, to

deal with these systematic judgmental errors, policymakers should consider other solutions besides relying on the current procedures. What is more, though the three examples (asylum refugee decisions, loan application reviews, MLB umpire pitch calls) provided by the authors are in diverse fields, they are all of a grant or deny choice. I extend their binary decisions to numeric scores in my project, as the numerics have greater generalizability.

2 Methodology

Following the pattern of the model by Chen, Moskowitz, and Shue (2016), which adapted from Rabin (2002), I utilize the model to test whether the agents experience negative autocorrelation and whether the gambler's fallacy is compatible with the negative autocorrelation.

I use the baseline model to test whether the current scores correlate with lagged scores, conditional on some control variables:

$$Y_{i,t} = \beta_0 + \beta_1 Y_{i,t-1} + Controls + \epsilon_{i,t}$$

where $Y_{i,t}$ is the score that decision maker i gives at time t, i.e., the score that judge i gives to the current show; β_1 captures the correlation between the current case and the previous case. If the merit of a competitor's performance is the only determinant of the score, given the order of evaluations are random, then β_1 will be 0 since the previous case should not predict the next case. If β_1 is negative, then it gives evidence of the negative autocorrelation in decisions.

I further construct a model of streaks to test whether the current change in scores is correlated with previous two trends:

$$\triangle Y_{i,t} = \beta_0 + \beta_1 \triangle Y_{i,t-1} + \beta_2 \triangle Y_{i,t-2} + Controls + \epsilon_{i,t}$$

where $\triangle Y_{it}$ is the change in score that judge i gives to the current show at t compared to the score they gave to the immediate prior performance at time $t-1$. To interpret the result, $\beta_1 < \beta_2 < 0$

implies the agent is under the gambler's fallacy.

In both models, the $controls$ is a set of variables, which controls for decision-maker heterogeneity using fixed effects and the moving average of previous decisions. The moving average regards the judges as one unit such that the differences between them are noises. Related control variables may include the panel average of the judge, indicators of previous drops by the judge, etc. But it is possible when one judge offers a high score, while another awards a low score that they cancel out. Then using the moving average alone would produce biases. Hence, I use the fixed effects model to control for judges separately. I also highlight the specification that additionally controls the performance fixed effects so that threats like the merits of the cases and the sequences themselves are negatively correlated are ruled out.

3 Data Selection

I use the observational data of the following two empirical examples to test the hypothesis.

3.1 Synchronized Swimming Competition

I use the data from five free routine preliminary competitions (Solo Free, Duet Free, Mixed Duet Free, Team Free, Free Combination) at the 2019 FINA World Aquatics Championships, which is the World Championships for aquatics sports that is held every two years by the International Swimming Federation (FINA) since 1973. In each of the preliminary competitions, the start order was randomized so that the dataset fits perfectly for the analysis. Each performance unit performed only once in the competition. To give evidence on the randomized order, I display the country names of the first three and last three performance units for both the 2019 and 2017 FINA Championships in Table 1.

Table 1: Synchronized Swimming: Nationalities of First and Last Three Performers

	2019 FINA		2017 FINA	
Competition Type	First three	Last three	First three	Last three
Solo Free	POL-Poland	ESP-Spain	TUR-Turkey	ISR-Israel
	CUB-Cuba	KOR-Korea	CHI-Chile	GBR-Great Britain
	UZB-Uzbekistan	LTU-Lithuania	CRC-Costa Rica	LIE-Liechtenstein
Duet Free	LIE-Liechtenstein	MAC-Macao	NED-Netherlands	ITA-Italy
	SVK-Slovakia	CAN-Canada	CHI-Chile	MAS-Malaysia
	KOR-Korea	SRB-Serbia	CRC-Costa Rica	UKR-Ukraine
Mixed Duet	AUS-Australia	JPN-Japan	GRE-Greece	BRA-Brazil
	UZB-Uzbekistan	USA-United States	USA-United States	RUS-Russia
	RUS-Russia	ESP-Spain	GER-Germany	CAN-Canada
Team Free	SUI-Switzerland	ESP-Spain	SGP-Singapore	KAZ-Kazakhstan
	JPN-Japan	POL-Poland	RSA-South Africa	USA-United States
	USA-United States	THA-Thailand	PRK-North Korea	ITA-Italy
Free Combination	JPN-Japan	ITA-Italy	N/A	N/A
	THA-Thailand	CHN-China		
	KAZ-Kazakhstan	BRA-Brazil		

Notes. This table presents the country names of the first three and last three performance units for the free routine preliminary competitions in 2019 and 2017 FINA Championships. For example, for the preliminary Solo Free in 2019 FINA, the first performer is from Poland, the second is from Cuba, the third is from Uzbekistan; the last is from Spain, the second from the last is Korean, the third from the last is Lithuanian.

I choose to examine the free routine since the other routine, technical, focused more on the required elements of the performance. The free routine depends on judges' subjectivity and hence, heuristics are more likely to play a role. There are three panels of judges: execution, artistic impression, and difficulty, with five judges in each panel, fifteen judges in total. The weighting factors for the panels are 30%, 40%, 30%, respectively. Each judge will award scores ranging from 0 to 10, with 0.1 increments, immediately after each performance. For every set of scores by a panel, the highest and lowest scores are dropped, and the weighted sum of the remaining scores minus any penalties are the total scores.

Table 2 summarizes the characteristics of my example of the synchronized swimming competition.

Table 2: Synchronized Swimming: Summary Statistics

	Mean	Std. Dev.	Median	Min	Max
Number of judges	75				
Number of performance units in Free Combination	15				
Number of performance units in Duet Free	45				
Number of performance units in Mixed Duet Free	11				
Number of performance units in Solo Free	32				
Number of performance units in Team Free	27				
Total number of performance units	130				
Total number of participating countries	47				
Enter next round indicator	0.46	0.498	0.00	0.00	1.00
Scores by individual judges	8.12	0.863	8.10	6.00	9.80
Scores by panels	9.01	1.590	8.73	6.33	9.77
Total scores	81.22	8.530	80.90	63.33	97.77

Notes. This table presents summary statistics of the synchronized swimming data used in the decision making analysis. The performance unit is 1 swimmer for Solo Free, 2 swimmers for Duet Free and Mixed Duet Free, 4 performers for Team Free, 10 performers for Free Combination.

In particular, to be eligible to judge at FINA events, the judges first need to attend FINA AS Development Schools Advanced Level and FINA AS Certification Schools consecutively to be certified. At least 60 hours per year of practical judging must be obtained after passing the test at each school. Then, all certified judges must pass an annual online exam through FINA online platform. Therefore, all judges in my example have met this rigid requirement and are well-experienced.

Control variables in the regressions of this example include a set of dummies for the number of drops over the past decisions (indicator for whether the immediate prior score by the judge is dropped; indicator for the number of drops out of the judge's previous two decisions excluding the current), and the judge's average score for previous three decisions excluding the current. These indicators control for the recent trends in judges' actions. I also include the average score of the panel of the judges for previous three decisions excluding the current, which control for the trends in evaluations at the panel-level. The model involves an additional control indicator for penalty of the current performance, which is a factor that may affect the impression of the performance on judges.

3.2 Hip Hop Dance Championship

The data composed of preliminary competitions from five divisions (Adult, Junior, Mega Crew, Mini Crew, Varsity) for both the World and the U.S. competition sections at the 2019 Hip Hop Dance Championships, which are held nationally and internationally every year. The start order was drawn randomly by computer before each of the preliminary competitions. Each performance group performs only once in the competition. There are only two panels of judges (performance and skill), with four judges in each. Each judge will award scores ranging from 0 to 5 after each show. Though there are prespecified criteria for the judges, the evaluations rely heavily on the subjective judgement of the judges. For every set of scores by a panel, the highest and lowest scores are dropped, and the sum of the panel averages minus any penalties are the total scores.

Table 3 displays the summary statistics of my example of hip hop dance championship.

Table 3: Hip Hop Dance: Summary Statistics

	Mean	Std. Dev.	Median	Min	Max
Number of judges	80(40)				
Number of performance groups in Adult	75(55)				
Number of performance groups in Junior	67(51)				
Number of performance groups in Mega Crew	70(57)				
Number of performance groups in Mini Crew	53(44)				
Number of performance groups in Varsity	91(64)				
Total number of performance groups	356(271)				
Total number of participating countries	45				
Enter next round indicator	0.21	0.406	0.00	0.00	1.00
Scores by individual judges	3.12	0.550	3.17	1.20	4.50
Scores by panels	3.12	0.484	3.16	1.45	4.32
Total scores	6.22	0.824	6.36	3.60	8.17

Notes. This table presents summary statistics of the hip hop data used in the decision making analysis. Numbers in the parenthesis are for the world competition sections (the rest unshown are for U.S. competition range).

Here, all judges in the Hip Hop Dance Championships are in Elite Judge Status, which requires several years of judging experience. These judges have participated in the official rules and regulations course, judged official Hip Hop Dance Championship events nationally and internationally, and successfully completed the Elite Judge Training Program. Generally, they can also be regarded

as well-experienced judges.

Control variables in the regressions of this example follow the ones in the previous synchronized swimming case.

4 Results

4.1 Synchronized Swimming: Results

Table 4: Synchronized Swimming: Individual-level Results

	Baseline					
	Dependent variable: score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag score	0.09070^{*}	-0.15718^{***}	0.51138^{***}	0.30150^{***}	-0.12607^{**}	-0.17107^{***}
	(0.04406)	(0.04666)	(0.11028)	(0.02839)	(0.04098)	(0.04347)
N	1725	1725	1725	1725	1725	1725
R^2	0.6765	0.79844	0.86902	0.84084	0.81163	0.80819
	Streaks					
	Dependent variable: differenced score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag differenced score	-0.59573^{***}	-0.71744^{***}	-0.37846^{***}	-0.66014^{***}	-0.70903^{***}	-0.71066^{***}
	(0.03550)	(0.02954)	(0.10322)	(0.03116)	(0.03215)	(0.03544)
β_2: Lag 2 differenced score	-0.29323^{***}	-0.26199^{***}	-0.20568^{*}	-0.33562^{***}	-0.35401^{***}	-0.37771^{***}
	(0.04046)	(0.04326)	(0.08300)	(0.03082)	(0.03175)	(0.03202)
N	1575	1575	1575	1575	1575	1575
R^2	0.83847	0.91118	0.32993	0.69271	0.90236	0.89964
Judge FE	No	Yes	Yes	Yes	Yes	Yes
Performance FE	No	No	Yes	No	No	No
Performer's Nationality FE	No	No	No	Yes	No	No
Competition FE	No	No	No	No	Yes	No
Judge Type FE	No	No	No	No	No	Yes

Notes. This table tests the hypotheses at individual-level. The baseline model tests whether the score to the performance by current competitor in the competition is related to the previous one. The streaks model tests whether the current change in score is related to the previous two trends. Column (2) controls for the judge fix effects. Column (3) controls for the performance fixed effects, Column (4) controls for the country of the competitors, Column (5) controls for competition type (Solo Free, Duet Free, Mixed Duet Free, Team Free, Free Combination), Column (6) controls for judge type (execution, artistic impression, difficulty), in addition to (2). All specifications include the following $controls$: indicator for whether the immediate prior score by the judge is dropped; indicator for the number of drops out of the judge's previous two decisions (excluding the current); the judge's average score for previous three decisions (excluding the current); the average score of the panel of the judge's for previous three decisions (excluding the current); indicator for penalty of the current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 5: Synchronized Swimming: Panel-level Results

	Baseline					
	Dependent variable: score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag score	0.17673	−0.00885	0.30436	0.10378	0.00233	−0.02459
	(0.20849)	(0.22220)	(0.16030)	(0.06698)	(0.23490)	(0.23555)
N	360	360	360	360	360	360
R^2	0.74908	0.86126	0.89101	0.88037	0.87487	0.86851
	Streaks					
	Dependent variable: differenced score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag	−0.53402***	−0.63547***	0.18213	−0.60977***	−0.58725***	−0.53270***
differenced score	(0.14434)	(0.14317)	(0.27725)	(0.11364)	(0.14203)	(0.13498)
β_2: Lag 2	−0.25485	−0.24972	−0.62202**	−0.36959**	−0.19458	−0.00583
differenced score	(0.13721)	(0.15343)	(0.21926)	(0.13202)	(0.17554)	(0.20471)
N	330	330	330	330	330	330
R^2	0.8664	0.92953	0.84239	0.94037	0.95165	0.94144
Judge FE	No	Yes	Yes	Yes	Yes	Yes
Performance FE	No	No	Yes	No	No	No
Performer's Nationality FE	No	No	No	Yes	No	No
Competition FE	No	No	No	No	Yes	No
Judge Type FE	No	No	No	No	No	Yes

Notes. This table tests the hypotheses at panel-level, i.e., using the sectional data of each panel after dropping scores. All specifications follow the ones in the previous table, the $controls$ include the panel's average score for previous three decisions (excluding the current); indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 6: Synchronized Swimming: Overall Results

Baseline				
	Dependent variable: score			
	(1)	(2)	(3)	(4)
β_1: Lag score	0.05427	0.04769	0.06212	0.25029**
	(0.07305)	(0.14114)	(0.07252)	(0.09292)
N	115	115	115	115
R^2	0.010767	0.01547	0.01161	0.08918
Streaks				
	Dependent variable: differenced score			
	(1)	(2)	(3)	(4)
β_1: Lag	-0.57336^{***}	-0.62777^{***}	-0.52768^{***}	-0.56667^{***}
differenced score	(0.02599)	(0.14024)	(0.11573)	(0.10432)
β_2: Lag 2	-0.35001^{***}	-0.44248^{***}	-0.29696^{***}	-0.25858^{***}
differenced score	(0.04394)	(0.03854)	(0.06969)	(0.04451)
N	105	105	105	105
R^2	0.27687	0.35098	0.29203	0.28416
Competition FE	No	Yes	Yes	Yes
Performance FE	No	No	Yes	No
Performer's Nationality FE	No	No	No	Yes

Notes. This table tests the hypotheses using the final scores of competitors. Column (2) controls for the Competition fixed effects. Column (3) controls for the performance fixed effects. Column (4) controls for the fixed effects of performer's nationality. The $controls$ include indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

For the synchronized swimming example, Table 4 - 6 display the results of autocorrelations at three levels (individual, panel, overall) with different specifications controlling for fixed effects. The three tables show negative and significant autocorrelations in judges' behavior in all specifications of the streaks model. That is, the judges are less likely to remain in the same direction given two of their immediate previous trends in scores. For the baseline model, however, there is no conclusive results that can account for all specifications, as the correlations diverse when we control for different fixed effects.

I extend the analysis further for Specification 3 of the streak model to see if the autocorrelations

last long for more subsequent decisions or they vanish quickly. As shown in the following Table 7 - 9, the autocorrelations between the score trends at t and up to at $t-4$ are negative and significant in all levels except for the panel-level, which implies that generally, the negative autocorrelations can persist long in a sequence of decisions. This is compatible with the conception of the gambler's fallacy.

Table 7: Synchronized Swimming: Individual-level Specification 3 Results

	Streaks		
	Dependent variable: differenced score		
	(1)	(2)	(3)
β_1: Lag differenced score	-0.37846^{***}	-0.43549^{***}	-0.55092^{***}
	(0.10322)	(0.11180)	(0.11729)
β_2: Lag 2 differenced score	-0.20568^{***}	-0.29784^{**}	-0.43529^{***}
	(0.08300)	(0.10643)	(0.11407)
β_3: Lag 3 differenced score		-0.26041^{**}	-0.42788^{***}
		(0.08805)	(0.10725)
β_4: Lag 4 differenced score			-0.36413^{***}
			(0.10949)
N	1575	1500	1425
R^2	0.32993	0.3891	0.47427

Notes. This table tests the hypotheses at individual-level. The streaks model tests whether the current change in score is related to the previous trends. It controls for the performance and judge fixed effects (as in Specification 3). Column (2) and Column (3) add more lagged variables respectively. All specifications include the following $controls$: indicator for whether the immediate prior score by the judge is dropped; indicator for the number of drops out of the judge's previous two decisions (excluding the current); the judge's average score for previous three decisions (excluding the current); the average score of the panel of the judge's for previous three decisions (excluding the current); indicator for penalty of the current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 8: Synchronized Swimming: Panel-level Specification 3 Results

	Streaks		
	Dependent variable: differenced score		
	(1)	(2)	(3)
β_1: Lag differenced score	0.18213	0.07736	0.11171
	(0.27725)	(0.25851)	(0.25123)
β_2: Lag 2 differenced score	-0.62202^{**}	-0.57015^{**}	-0.77951^{***}
	(0.21926)	(0.16854)	(0.17318)
β_3: Lag 3 differenced score		-0.28696	-0.14596
		(0.15112)	(0.13581)
β_4: Lag 4 differenced score			-0.88860^{**}
			(0.31921)
N	330	315	300
R^2	0.84239	0.85156	0.88564

Notes. This table tests the hypotheses at panel-level, i.e., using the sectional data of each panel after dropping scores. It controls for the performance and judge fixed effects (as in Specification 3). All specifications follow the ones in the previous table, the $controls$ include the panel's average score for previous three decisions (excluding the current); indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 9: Synchronized Swimming: Overall Specification 3 Results

	Streaks		
	Dependent variable: differenced score		
	(1)	(2)	(3)
β_1: Lag differenced score	-0.52768^{***}	-0.65291^{***}	-0.69531^{***}
	(0.11573)	(0.09517)	(0.08983)
β_2: Lag 2 differenced score	-0.29696^{***}	-0.44079^{***}	-0.54710^{***}
	(0.06969)	(0.07453)	(0.11026)
β_3: Lag 3 differenced score		-0.35465^{***}	-0.50118^{***}
		(0.05874)	(0.07654)
β_4: Lag 4 differenced score			-0.22077^{***}
			(0.05051)
N	105	100	95
R^2	0.29203	0.39123	0.41798

Notes. This table tests the hypotheses using the final scores of competitors. It controls for the performance and judge fixed effects (as in Specification 3). All specifications follow the ones in the previous table. The $controls$ include indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

4.1.1 Robustness

Table 10: Synchronized Swimming: Technical Results

Baseline		
	Dependent variable: score	
	Individual × Technical	Panel × Technical
β_1: Lag score	-0.10223^{*}	0.13436^{***}
	(0.04445)	(0.03511)
N	1050	215
R^2	0.30759	0.15596
Streaks		
	Dependent variable: differenced score	
	Individual × Technical	Panel × Technical
β_1: Lag differenced score	-0.53771^{***}	-0.15057
	(0.02630)	(0.22973)
β_2: Lag 2 differenced score	-0.28943^{**}	-0.36217
	(0.03216)	(0.31227)
N	1000	205
R^2	0.45991	0.13001

Notes. This table tests the hypotheses at individual and panel level using the scores from duet technical in 2019 FINA. All specifications in the table controls for the judge fixed effects and performance fixed effects. The $controls$ include the variables specified in previous free routine analyses respectively. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

I replicate the analysis using the preliminary duet technical scores to check for robustness. The results in Table 10 show that there exist negative autocorrelations even in the technical routine that is more rigid in the rules. With procedures of dropping extreme scores and taking average scores, the autocorrelations turn insignificant or even positive. The findings suggest two possible cases: judges may be heavily influenced by the gambler's fallacy even with rigorous criteria and rules to obey; or, it is possible that scoring in the technical routine also involves subjective judgments where judges rely on heuristics.

4.2 Hip Hop Dance: Results

Table 11: Hip Hop Dance: Individual-level Results

	Baseline					
	Dependent variable: score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag score	-0.00738	-0.03697	0.07006^{***}	0.08472^{***}	-0.04551	-0.04467
	(0.03341)	(0.03495)	(0.01901)	(0.02291)	(0.03468)	(0.03592)
N	2608	2608	2608	2608	2608	2608
R^2	0.31012	0.32714	0.3515	0.354	0.34748	0.34201
	Streaks					
	Dependent variable: differenced score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag differenced score	-0.66765^{***}	-0.68965^{***}	-0.66964^{***}	-0.68047^{***}	-0.67116^{***}	-0.68293^{***}
	(0.02293)	(0.02314)	(0.02072)	(0.02135)	(0.02514)	(0.02278)
β_2: Lag 2 differenced score	-0.28360^{***}	-0.29599^{***}	-0.32424^{***}	-0.32862^{***}	-0.26507^{***}	-0.26845^{***}
	(0.02509)	(0.02385)	(0.01882)	(0.01839)	(0.02281)	(0.02499)
N	2448	2448	2448	2448	2448	2448
R^2	0.67082	0.69031	0.62662	0.61665	0.69225	0.69591
Judge FE	No	Yes	Yes	Yes	Yes	Yes
Performance FE	No	No	Yes	No	No	No
Competitor region FE	No	No	No	Yes	No	No
Competition type FE	No	No	No	No	Yes	No
Judge type FE	No	No	No	No	No	Yes

Notes. This table tests the hypotheses at individual-level. The baseline model tests whether the score to the performance by current competitor in the competition is related to the previous one. The streaks model tests whether the current change in score is related to the previous two trends. Column (2) controls for the judge fix effects. Column (3) controls for the competitor fix effects, Column (4) controls for the region of the competitors, Column (5) controls for competition type (Adult, Junior, Mega Crew, Mini Crew, Varsity), Column (6) controls for judge type (performance, skill), in addition to (2). All specifications include the following $controls$: indicator for whether the immediate prior score by the judge is dropped; indicator for the number of drops out of the judge's previous two decisions (excluding the current); the judge's average score for previous three decisions (excluding the current); indicator for penalty of the current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 12: Hip Hop Dance: Panel-level Results

	Baseline					
	Dependent variable: score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag score	-0.12595	-0.09732^{*}	0.083255	0.00669	-0.10084^{**}	-0.11964^{**}
	(0.11020)	(0.03930)	(0.04526)	(0.02704)	(0.03586)	(0.03980)
N	672	672	672	672	672	672
R^2	0.65275	0.01504	0.72331	0.00672	0.016464	0.021919
	Streaks					
	Dependent variable: differenced score					
	(1)	(2)	(3)	(4)	(5)	(6)
β_1: Lag differenced score	-0.75768^{***}	-0.75719^{***}	-0.68371^{***}	-0.70252^{***}	-0.71730^{***}	-0.74825^{***}
	(0.08126)	(0.04149)	(0.03520)	(0.03078)	(0.03919)	(0.03558)
β_2: Lag 2 differenced score	-0.31664^{***}	-0.32098^{***}	-0.32767^{***}	-0.34941^{***}	-0.30381^{***}	-0.32850^{***}
	(0.04993)	(0.04256)	(0.03940)	(0.03493)	(0.04055)	(0.03288)
N	632	632	632	632	632	632
R^2	0.83586	0.39549	0.40544	0.43715	0.37429	0.39311
Judge FE	No	Yes	Yes	Yes	Yes	Yes
Performance FE	No	No	Yes	No	No	No
Competitor region FE	No	No	No	Yes	No	No
Competition type FE	No	No	No	No	Yes	No
Judge type FE	No	No	No	No	No	Yes

Notes. This table tests the hypotheses at panel-level, i.e., using the data of the average of each panel after dropping scores. All specification follows the ones in the previous table, the variable $controls$ include indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 13: Hip Hop Dance: Overall Results

	Baseline				
	Dependent variable: score				
	(1)	(2)	(3)	(4)	(5)
β_1: Lag score	-0.18550^{*}	-0.12940^{*}	0.02693	0.07909^{*}	-0.09780^{*}
	(0.08654)	(0.06361)	(0.04592)	(0.03253)	(0.04826)
N	248	336	336	336	336
R^2	0.02943	0.01650	0.00093	0.00701	0.01008
	Streaks				
	Dependent variable: differenced score				
	(1)	(2)	(3)	(4)	(5)
β_1: Lag differenced score	-0.73816^{***}	-0.78544^{***}	-0.62547^{***}	-0.59470^{***}	-0.71550^{***}
	(0.06394)	(0.06413)	(0.04940)	(0.04998)	(0.03308)
β_2: Lag 2 differenced score	-0.26765^{***}	-0.30392^{***}	-0.36867^{***}	-0.32571^{***}	-0.24526^{***}
	(0.07316)	(0.06772)	(0.06661)	(0.05916)	(0.06324)
N	316	316	316	316	316
R^2	0.38427	0.41538	0.32589	0.29754	0.3785
Competition FE	No	Yes	Yes	Yes	Yes
Performance FE	No	No	Yes	No	No
Competitor region FE	No	No	No	Yes	No
Competition type FE	No	No	No	No	Yes

Notes. This table tests the hypotheses using the final scores of competitors. Column (2) controls for the Competition fixed effects. Column (3) controls for the fixed effects of performer. Column (4) controls for the fixed effects of performer's region. Column (5) controls for the fixed effects of type of competition. The $controls$ only include indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

When it comes to the hip hop dance example, Table 11 - 13 show the results which agree with the findings of the previous example. The judges in non-sport hip hop dance competitions are more likely to adjust their trends to offer high scores following the downward trends they experience.

Extending the Specification 3 of the streak model, we can find from Table 14 - 16 that the autocorrelations between the score trends at t and up to at $t-4$ are negative and significant in all levels, which implies that the findings from the synchronized swimming example also hold. The negative and significant autocorrelations suggest that the behavior of the judges is compatible with

the interpretation by the gambler's fallacy.

Table 14: Hip Hop Dance: Individual-level Specification 3 Results

	Streaks		
	Dependent variable: differenced score		
	(1)	(2)	(3)
β_1: Lag differenced score	-0.66964^{***}	-0.76141^{***}	-0.81186^{***}
	(0.02072)	(0.02383)	(0.02227)
β_2: Lag 2 differenced score	-0.32424^{***}	-0.49474^{***}	-0.60778^{***}
	(0.01882)	(0.02190)	(0.02426)
β_3: Lag 3 differenced score		-0.25662^{***}	-0.41875^{***}
		(0.02122)	(0.02434)
β_4: Lag 4 differenced score			-0.20630^{***}
			(0.01810)
N	2448	2368	2288
R^2	0.62662	0.66685	0.69294

Notes. This table tests the hypotheses at individual-level. The streaks model tests whether the current change in score is related to the previous trends. It controls for the performance and judge fixed effects (as in Specification 3). Column (2) and Column (3) add more lagged variables respectively. All specifications include the following $controls$: indicator for whether the immediate prior score by the judge is dropped; indicator for the number of drops out of the judge's previous two decisions (excluding the current); the judge's average score for previous three decisions (excluding the current); indicator for penalty of the current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 15: Hip Hop Dance: Panel-level Specification 3 Results

	Streaks		
	Dependent variable: differenced score		
	(1)	(2)	(3)
β_1: Lag differenced score	-0.68371^{***}	-0.76223^{***}	-0.77981^{***}
	(0.03520)	(0.02788)	(0.03154)
β_2: Lag 2 differenced score	-0.32767^{***}	-0.50113^{***}	-0.55786^{***}
	(0.03940)	(0.04011)	(0.04807)
β_3: Lag 3 differenced score		-0.26736^{***}	-0.37928^{***}
		(0.04520)	(0.05918)
β_4: Lag 4 differenced score			-0.15088^{***}
			(0.03881)
N	632	612	592
R^2	0.40544	0.45763	0.4718

Notes. This table tests the hypotheses at panel-level, i.e., using the sectional data of each panel after dropping scores. It controls for the performance and judge fixed effects (as in Specification 3). All specifications follow the ones in the previous table, the $controls$ only include indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

Table 16: Hip Hop Dance: Overall Specification 3 Results

	Streaks		
	Dependent variable: differenced score		
	(1)	(2)	(3)
β_1: Lag	-0.62547^{***}	-0.72672^{***}	-0.79646^{***}
differenced score	(0.04940)	(0.04550)	(0.04805)
β_2: Lag 2	-0.36867^{***}	-0.53062^{***}	-0.68631^{***}
differenced score	(0.06661)	(0.06586)	(0.06249)
β_3: Lag 3		-0.24829^{***}	-0.46465^{***}
differenced score		(0.06454)	(0.07395)
β_4: Lag 4			-0.29925^{***}
differenced score			(0.04859)
N	316	306	296
R^2	0.32589	0.36527	0.41768

Notes. This table tests the hypotheses using the final scores of competitors. It controls for the performance and judge fixed effects (as in Specification 3). All specifications follow the ones in the previous table. The $controls$ include indicator variables for penalty of current performance. Robust standard errors in brackets. *, **, *** indicate significance at the 10%, 5%, 1% levels, respectively.

5 Discussion

5.1 Threats and Limitations

The greatest threat to the current research design is the internal validity, as I need to clean the data before analyzing the behavior of the judges. With the essential assumption that the order of performances is randomized by the rigorous rule, and the scoring procedure is under strict supervision, my analysis results are credible to some extent. With the additional control variables mentioned above, the power of the validity strengthens further. The rigorous requirements for a judge to be qualified in both examples ensures that judges are well-experienced. Combined with the results, the market experience that may help mitigate anomalies in the paper by List (2011) is not applicable in my case. The conclusion of the persistence of judgmental biases in competitions is concrete.

There are some alternative explanations for the negative autocorrelation. Firstly, it may include learning effects. But just as the market experience mentioned above, it is unlikely given the ex-

perienced judges. Quotas on scores are also less likely since judges claim to be unrestricted on the scores they are giving. Most importantly, the sequential contrast effects (SCE) suggested by prior literature could also account for the results. Theoretically, the main difference between the gambler's fallacy and SCE is that, under the gambler's fallacy, the subjects predict the quality of the upcoming case even before the assessment. While with SCE, the comparison is made only after seeing the performance. It is difficult to distinguish as the procedures are unobservable with naturally occurring data. Even with the model that controls the lagged quality suggested by Chen, Moskowitz, and Shue (2016), SCE could not be fully rejected, as the unobserved quality measure may better capture the true quality.

One potential limitation of the current study is that the heterogeneity analysis is absent due to the limited information about judges in the examples. If I had data about the basic characteristics of the judges, like gender, nationality, age, years as a judge, education level, etc., it would be interesting to analyze whether these covariates play a role when judges make sequential decisions.

Testing the hypothesis in a laboratory setting would avoid these issues, as threats to the internal validity are under control. The predictions of the quality of the subsequent performances before seeing can be obtained from the subjects to rule out SCE. An additional survey could be conducted to determine if subjects auto correlate due to these alternative reasons. Abundant information of subjects' characteristics can be gathered for heterogeneity analysis. However, this is difficult to implement as it requires a large amount of performers and judges.

5.2 Implications

For the results shown in Table 4 - 16, there are negative and significant autocorrelations among the sequences of decisions, which persist even after procedures such as dropping scores and taking average scores. Policymakers should therefore consider other modifications before or during the evaluation process to assure the fairness of the judgments. One possible approach may be correcting judges' erroneous beliefs before awarding the scores. Follow-up investigations may turn to experimentation to see the effectiveness of such adjustments. For example, if we can gather many

judges to evaluate a randomized sequence of performances in a given field, just as the experiment by Bursztyn, González, and Yanagizawa-Drott (2020), we can assign half of the subjects to the treatment group where they are provided the information about the gambler's fallacy before the evaluations. By analyzing using the difference-in-differences method, we can conclude whether information provision is an effective alternative way of improving the evaluation impartiality. If so, then the information of the gambler's fallacy effect may be introduced in the judges' training program and highlighted at the beginning of every competition.

Policymakers can also try to mitigate the effect by adjusting the scoring procedure. For example, if each judge evaluates the performance in a different randomized order, then the bias caused by the gambler's fallacy is distributed equally on scores given by the judge. It might be a good way to make the competition fair, as there is no prespecified "sequence" in this case, and therefore, no sequential judgmental bias.

On the other hand, if the biases cannot be easily corrected, and multiple judges perform as a single judge in the competition, then from an economic perspective, the competition committee may consider reducing the size of the judging panel to save the budget.

Beyond the competition context of my study, the negative autocorrelation under the gambler's fallacy may be found in other settings such as, in essay grading, language evaluation examinations, program admissions, drug review, and financial auditing. The results of my research may be suggestive for these examples that erroneous belief may lead to biased sequential judgments.

6 Conclusion

Using two empirical examples of synchronized swimming competition and hip hop dance championship, I investigated the behavior of judges in competitions under the gambler's fallacy. The results show that there exists negative autocorrelation that is independent of the performers' merits, and is compatible with the interpretation of the gambler's fallacy, where individuals underestimate the occurrence probability of the sequential streaks. The findings suggest that such sequential

judgment biases last long and persist with the current scoring procedure design. The spontaneous adjustments by judges backfire the goal of impartiality in competitions, which is a question left for the related policymakers. Further improvements such as correcting judges' erroneous beliefs and adjusting the rules may be needed to mitigate the effect.

References

Bruin, W. B. d. (2005). Save the last dance for me: Unwanted serial position effects injury evaluations. *Acta Psychologica 118*(3), 245–260.

Bruin, W. B. d. (2006). Save the last dance ii: Unwanted serial position effects in figure skating judgments. *Acta Psychologica 123*(3), 299–311.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American Economic Review 110*(10), 2997–3029.

Chen, D. L., T. J. Moskowitz, and K. Shue (2016, August). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics 131*(3), 1181–1242.

Damisch, L., T. Mussweiler, and H. Plessner (2006). Olympic medals as fruits of comparison? assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied 12*(3), 166–178.

Flôres, J. R. G. and V. A. Ginsburgh (1996). The queen elisabeth musical competition: How fair is the final ranking? *Journal of the Royal Statistical Society 45*(1), 97–104.

Kramer, R. S. S. (2017). Sequential effects in olympic synchronized diving scores. *Royal Society Open Science 4*(160812), 1–9.

List, J. A. (2011). Does market experience eliminate market anomalies? the case of exogenous market experience. *American Economic Review 101*(3), 313–317.

Mantonakis, A., P. Rodero, I. Lesschaeve, and R. Hastie (2009). Order in choice: Effects of serial position on preferences. *Psychological Science 20*(11), 1309–1312.

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

Rabin, M. (2002, August). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics 117*(3), 775–816.

Tversky, A. and D. Kahneman (1974, September). Judgment under uncertainty: Heuristics and biases. *Science, New Series 185*(4157), 1124–1131.