PEER REVIEW EXHIBITS THE GAMBLER'S FALLACY

Scott Carrell[a], David Figlio[b]*, Lester Lusher[c]

[a]University of California, Davis, NBER, and IZA

[b]Northwestern University, NBER, and IZA

[c]University of Hawaii at Manoa and IZA

This draft: April 19, 2020

Abstract

Using data on reviewer recommendations for submitted manuscripts from a top field journal in economics, where we observe multiple reviewers per paper and multiple papers per reviewer, we find evidence of a negative autocorrelation in reviews: A reviewer is significantly more likely to give a negative recommendation on a manuscript if their recommendation on the manuscript they most recently reviewed was positive. This phenomenon cannot be explained by potential differences in paper quality assigned across reviewers. This evidence is consistent with the gambler's fallacy - reviewers underestimate the probability of streaks occurring by chance, leading to negatively autocorrelated recommendations. Proxied with various measures, reviewers of greater expertise are less susceptible to this issue. Furthermore, female reviewers exhibit zero negative autocorrelation. We find no evidence that particular types of authors were harmed from these biases.

*Corresponding author: David Figlio, Northwestern University, School of Education and Social Policy, 2120 Campus Drive, Evanston, IL 60208 United States. Email: figlio@northwestern.edu; Scott Carrell, Email: secarrell@ucdavis.edu; Lester Lusher, Email: lrlusher@hawaii.edu

1 Introduction

Academia today is often regarded as ultracompetitive, with publishing in top academic journals becoming increasingly important in determining tenure and promotion [1–4]. A key tenet in the process of evaluating scientific research for publication is peer review [5]. Ideally, reviewers objectively evaluate the merits of each article when making recommendations for publication. That is, each article is treated as an independent draw from the distribution of research and factors other than the paper's quality are not taken into account. Prior research has shown this may not be the case, with biases in the review process shown by both gender [6, 7] and stature in the profession [8].

In addition, with the number of papers being submitted for peer review growing each year and reviewers becoming overwhelmed with requests to serve as peer reviewers [9–12], there may be circumstances in which reviewers are more likely to make Type I or Type II errors. Research from psychology and behavioral economics has demonstrated that when making objective judgments of independent events, people will often fall victim to a negative autocorrelation in their decision making due to the "law of small numbers" or the "gambler's fallacy" [13–15]. For example, studies have shown that when people are asked to generate a hypothetical sequence of random coin tosses, their sequences tend to contain more alternations than would be expected by chance [16, 17]. Decision making in high-stakes settings also appear to be susceptible to such biases, including in refugee asylum court rulings, loan application reviews, Major League Baseball umpire pitch calls, and lottery player decisions [18, 19].

We investigate whether reviewers of academic research are also susceptible to this "gambler's fallacy" by displaying a negative autocorrelation in their recommendations for publication. To do so, we make use of a unique dataset of reviewers' interactions with a leading field journal in economics that practices single-blind reviewing. Each paper is reviewed by multiple reviewers, and since reviewers review for the journal with relative frequency, we observe multiple papers per reviewer. Both facts permit the ability to study the "gambler's fallacy" in academic peer review.

Importantly, by using papers with multiple reviewers, we are able to compare the recommendations from two (or more) reviewers of the same paper with differing prior review histories. Specifically, our data allow us to estimate regression models while including paper fixed effects which control for all paper-specific factors such as the paper's quality or suitability for publication [20, 21]. At its core, our analysis simply compares how two (or more) otherwise identical reviewers of the same paper differ based on their prior review recommendation.

2 Results

Table 1 presents summary statistics for our analytic sample. We observe 286 reviewers across 431 papers that meet our criteria for inclusion (the paper made it past the initial editor review threshold and was reviewed by two or more reviewers, each with three or more observed reviews). Roughly a third of reviewers are female, three-quarters are North American, and 85% are White. Over half of the sample of reviewers received their PhD from a top 10 PhD program (as ranked by US News). Turning to the unit of observation (reviewer-paper level, n=929), a little over 40% of reviewers positively evaluate their assigned manuscript. Reviewers also have the ability to give a strong positive recommendation, such as "Publish as is" or "Only minor revisions required"; around 7% of recommendations in the analytic sample constitute a strong positive recommendation. Editors accept approximately 20% of manuscripts in our sample.[1]

Our results also consider tests by the expertise/experience of the reviewer, since more experienced reviewers are potentially less likely to be susceptible to a negative autocorrelation [18]. For instance, we consider "senior" reviewers (8+ years removed from their PhD) who constitute around 60% of the sample. We also consider affiliation with the National Bureau of Economic Research (NBER), an institution which is typically regarded as an exclusive group of top economics researchers. Last, we consider whether the reviewer had previously published in one of five "top general interest" economics journals, another signal of reviewer prestige. Roughly 40% of reviewers are NBER affiliated and 45% of reviewers had previously published in a top general interest journal at the time the review was conducted.

Table 3 presents our main results. In column (1), where we do not condition on paper fixed effects, we find that reviewers are 12 percentage points less likely to give a positive recommendation if their prior recommendation was positive. When we include paper fixed effects in column (2), the estimate shifts only slightly to 11.5 percentage points, evidence there is no endogenous assignment of manuscripts (e.g., manuscript's quality) from editors to reviewers based on the reviewer's prior recommendation. In columns (3) and (4) we again find a negative autocorrelation based on whether the reviewer made a strong positive prior recommendation. In columns (5) and (6) we find that the editor's decision on the reviewer's prior manuscript also influences the reviewer's recommendation: If the editor ultimately accepted (rejected) the reviewer's prior manuscript, then the reviewer is 15.1 percentage points less (more) likely to positively evaluate their current manuscript.

Figure 1 investigates whether "expert" reviewers are less susceptible to a negative autocorrelation. To do so, we estimate our full model in column (2) from Table 3 while interacting our indicator for the re-

[1] Our sample is conditional on the paper receiving at least two reviews. Since it is common for editors to reject papers with zero or one review, the overall acceptance rates at this journal is less than 5%.

viewer's prior recommendation with various indicators of their experience at the time of the current review. This allows us to separately identify whether there are stronger negative autocorrelations across types of reviewers. For instance, in the first pairing we find that reviewers affiliated with the NBER are less susceptible to a negative autocorrelation (3.8 percentage point drop in likelihood of a positive review if prior review was positive) than non-NBER reviewers (17.6 percentage point drop). Similar patterns hold when we consider whether the reviewer published in a top general journal, was eight or more years removed from their PhD, and whether they ex post conducted five or more reviews for the journal. In fact, for all four of our measures of expertise, we find that expert reviewers do not display statistically significant biases in their recommendations.

In Figure 1, we also interact our indicator for the reviewer's prior recommendation with whether the reviewer is female. We find that the negative autocorrelation is driven exclusively by male reviewers (17.9 percentage point drop in likelihood of a positive review if prior review was positive), whereas female reviewers are completely unsusceptible to the gambler's fallacy. Consequently, this reinforces findings from prior studies, albeit in very different settings, that women tend to be the "fairer" sex [22–24].

Finally, in Figure 2 we investigate whether these fluctuations in reviewer recommendations differentially harm or benefit papers written by different types of authors. Much like Figure 1, we do this by interacting the reviewer's prior recommendation with author characteristics. Overall, we find that all authors, regardless of stature, experience or gender equally fall victim to reviewer negative autocorrelation biases.

3 Conclusion and Discussion

Peer review, the process by which one's work is evaluated by a set of peer scholars, is ubiquitous in academia. Hiring, promotion, and salary decisions are largely determined by one's publication success. As such, researchers have long been concerned with potential pitfalls in this process. Indeed, studies have shown that biases against certain types of papers or authors exist [6–8].

In this paper, we identify a bias in decision making found in other settings but not yet explored in peer review – a bias generated from the "gambler's fallacy" [15]. When tasked with generating predictions or judgments on objective and independent events, studies have shown that individual decisions tend to contain more fluctuations than one would expect by statistical chance, resulting in a negative autocorrelation across choices [16–19]. In our setting, we find that reviewers are significantly more likely to recommend a rejection (acceptance) if their recommendation on their prior manuscript was positive (negative). Importantly, this result holds even when accounting for potential differences in paper quality across referees by estimating models with paper fixed effects. Our results also suggest that expert reviewers and female reviewers are less susceptible to this bias and these biases equally affect all types of authors.

Much like prior settings, this negative correlation may arise from reviewers feeling they are “due” for an alternate outcome based on their recent prior review history. Reinforcing this is our finding that reviewer recommendations deviate from the editor’s final decision on the reviewer’s prior assignment (i.e. “my prior assigned paper was accepted by the editor so I’m due for a rejection” or vice versa). But especially in fields (such as economics) where a tenurable faculty member may only have a small number of publications, the randomness introduced by peer reviewers exhibiting the “gamblers fallacy” could seriously affect careers, for better or for worse.

References

[1] Svein Kyvik. Changing trends in publishing behaviour among university faculty, 1980-2000. *Scientometrics*, 58(1):35–48, 2003.

[2] Lesley A Schimanski and Juan Pablo Alperin. The evaluation of scholarship in academic promotion and tenure processes: Past, present, and future. *F1000Research*, 7, 2018.

[3] John Robert Warren. How much do you have to publish to get a job in a top sociology department? or to get tenure? trends over a generation. *Sociological Science*, 6:172–196, 2019.

[4] James J Heckman and Sidharth Moktan. Publishing and promotion in economics: The tyranny of the top five. *Journal of Economic Literature*, forthcoming.

[5] Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. Peer review in scientific publications: benefits, critiques, & a survival guide. *EJIFCC*, 25(3):227, 2014.

[6] Julian Kolev, Yuly Fuentes-Medel, and Fiona Murray. Is blinded review enough? how gendered outcomes arise even under anonymous evaluation. Technical report, National Bureau of Economic Research, 2019.

[7] David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. Are referees and editors in economics gender neutral? Technical report, National Bureau of Economic Research, 2019.

[8] Scott Carrell, David Figlio, and Lester Lusher. Clubs and networks in economics reviewing. Technical report, 2020.

[9] Michael E Hochberg, Jonathan M Chase, Nicholas J Gotelli, Alan Hastings, and Shahid Naeem. The tragedy of the reviewer commons. *Ecology Letters*, 12(1):2–4, 2009.

[10] Bruno S Frey, Reiner Eichenberger, and René L Frey. Editorial ruminations: publishing kyklos. *Kyklos*, 62(2):151–160, 2009.

[11] Alex Csiszar. Peer review: Troubled from the start. *Nature News*, 532(7599):306, 2016.

[12] Rafael DAndrea and James P ODwyer. Can editors save peer review from peer reviewers? *PloS one*, 12(10):e0186111, 2017.

[13] GS Tune. Response preferences: A review of some relevant literature. *Psychological bulletin*, 61(4): 286, 1964.

[14] Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76 (2):105, 1971.

[15] Matthew Rabin and Dimitri Vayanos. The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2):730–778, 2010.

[16] Maya Bar-Hillel and Willem A Wagenaar. The perception of randomness. *Advances in applied mathematics*, 12(4):428–454, 1991.

[17] Amnon Rapoport and David V Budescu. Randomization in individual choice behavior. *Psychological Review*, 104(3):603, 1997.

[18] Daniel L Chen, Tobias J Moskowitz, and Kelly Shue. Decision making under the gamblers fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131(3):1181–1242, 2016.

[19] Sigrid Suetens, Claus B Galbo-Jørgensen, and Jean-Robert Tyran. Predicting lotto numbers: a natural experiment on the gambler's fallacy and the hot-hand fallacy. *Journal of the European Economic Association*, 14(3):584–607, 2016.

[20] Pietro Balestra and Marc Nerlove. Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica (pre-1986)*, 34(3):585, 1966.

[21] Sergio Correia. Reghdfe: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects. 2018.

[22] James Andreoni and Lise Vesterlund. Which is the fair sex? gender differences in altruism. *The Quarterly Journal of Economics*, 116(1):293–312, 2001.

[23] David Dollar, Raymond Fisman, and Roberta Gatti. Are women really the fairer sex? corruption and women in government. *Journal of Economic Behavior & Organization*, 46(4):423–429, 2001.

[24] Paul S Herrnson and Jennifer C Lucas. The fairer sex? gender and negative campaigning in us elections. *American Politics Research*, 34(1):69–94, 2006.

[25] Aboozar Hadavand, Daniel S Hamermesh, and Wesley W Wilson. Is scholarly refereeing productive (at the margin)? Technical report, National Bureau of Economic Research, 2020.

4 Tables and Figures

Table 1: Summary statistics

	Reviewer-level	
	Mean	SD
Female	0.37	0.48
Nationality:		
-North American	0.76	0.43
-Asian	0.03	0.17
-European	0.16	0.36
-Other/missing	0.06	0.24
Race:		
-White	0.85	0.35
-Asian	0.08	0.27
-Other/missing	0.07	0.25
Institution of PhD (US News):		
-Ranked top 10	0.51	0.50
-Ranked 11-30	0.28	0.45
-Ranked 30-50	0.04	0.19
-Ranked 50+ / missing	0.17	0.38
Year receive PhD	2005.21	6.53
Unknown PhD year	0.05	0.22
Observations	286	
	Paper-level	
	Mean	SD
Number of reviewers	2.16	0.38
No NBER authors	0.80	0.40
No authors publish in top general interest	0.69	0.46
All junior authors	0.43	0.50
All male authors	0.45	0.50
Observations	431	
	Reviewer-paper level	
	Mean	SD
Outcome: Reviewer positive review	0.41	0.49
Positive prior review	0.42	0.49
(Conditional) accept prior review	0.07	0.25
Editor accept prior paper	0.21	0.41
Reviewer NBER affiliated	0.42	0.49
Reviewer published in top general interest	0.46	0.50
Reviewer senior ($>$ 7 years since PhD)	0.62	0.49
Observations	929	

Notes: The NBER (National Bureau of Economic Research) is an organization that disseminates research findings among academics affiliated with the organization. The NBER is typically regarded as an exclusive group of the top academics in the economics profession. An author/reviewer is considered published in a top general interest journal if they published in at least one of the *Quarterly Journal of Economics*, *American Economic Review* (not including *Papers & Proceedings*), *Review of Economic Studies*, *Econometrica*, and *Journal of Political Economy*.

Table 2: Autocorrelation in reviewer recommendations

	Lagged reviewer reports					
	(1)	(2)	(3)	(4)	(5)	(6)
Outcome: Reviewer positive review						
Positive prior review	-0.120***	-0.115**				
	(0.037)	(0.048)				
(Conditional) accept prior review			-0.109*	-0.150*		
			(0.063)	(0.081)		
Editor accept prior paper					-0.112***	-0.151**
					(0.042)	(0.063)
Observations	929	929	929	929	929	929
Reviewer FE	X	X	X	X	X	X
Paper FE		X		X		X

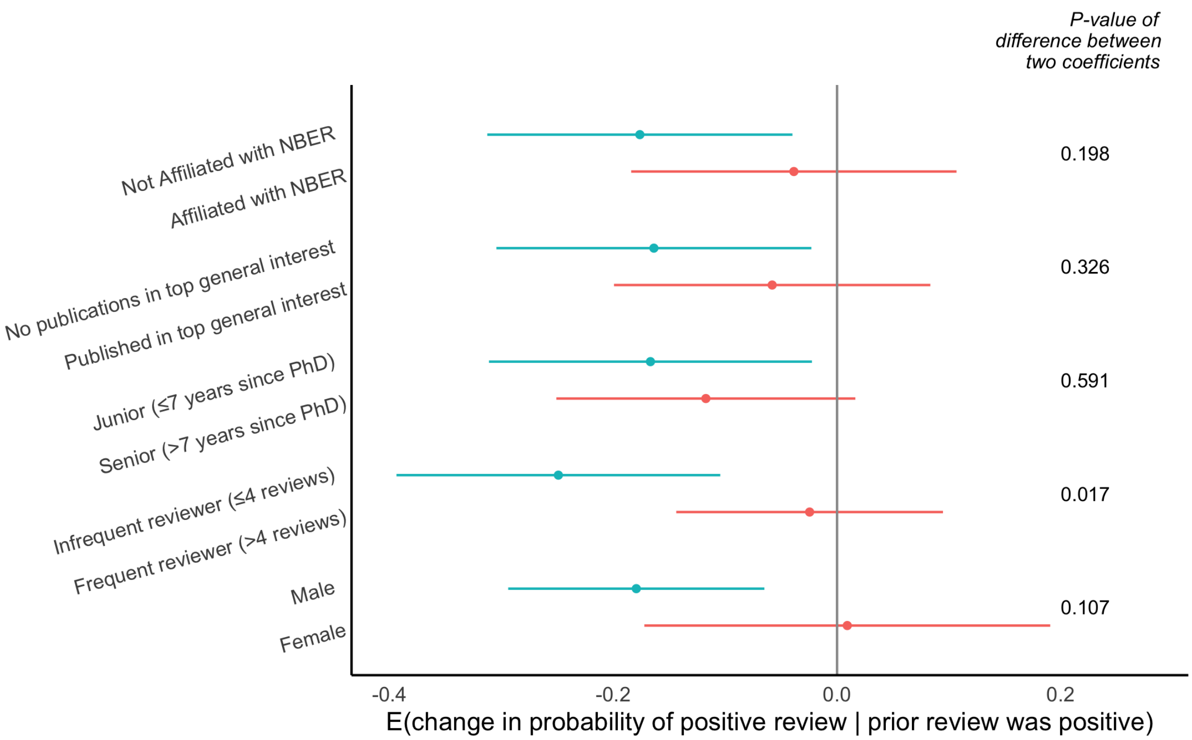
Notes: Each regression includes dummies for number of reviews done prior to observation and a linear control for number of months since the reviewer's last review. Standard errors clustered at the reviewer level. One, two, and three asterisks indicate statistical significance at the 10, 5, and 1 percent levels, respectively.

Table 3: Autocorrelation in reviewer recommendations

	Lagged reviewer reports					
	(1)	(2)	(3)	(4)	(5)	(6)
Outcome: Reviewer positive review						
Positive prior review	0.125***	0.118*				
	(0.045)	(0.065)				
(Conditional) accept prior review			0.101	0.071		
			(0.065)	(0.090)		
Editor accept prior paper					-0.049	-0.111*
					(0.040)	(0.058)
Observations	927	925	927	925	927	925
Reviewer avg. RR rate	X	X	X	X	X	X
Paper FE		X		X		X

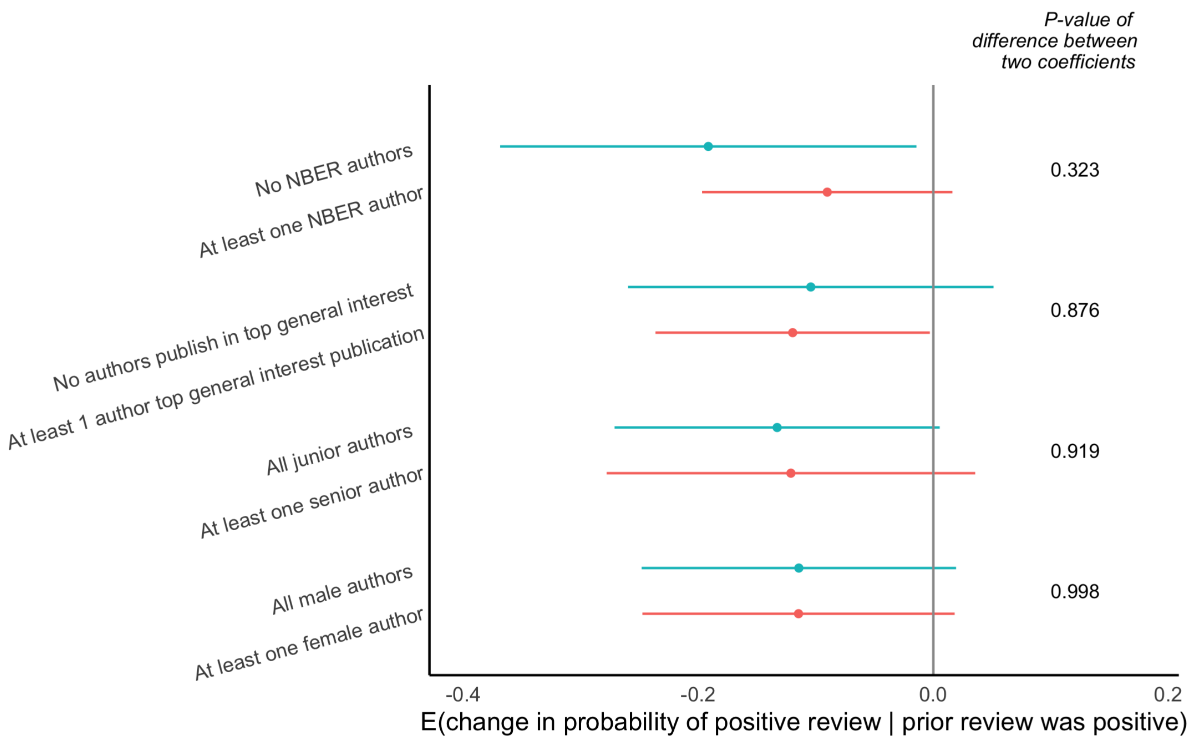
Notes: Each regression includes dummies for number of reviews done prior to observation and a linear control for number of months since the reviewer's last review. Standard errors clustered at the reviewer level. One, two, and three asterisks indicate statistical significance at the 10, 5, and 1 percent levels, respectively.

Figure 1: Autocorrelation in reviewer recommendations - reviewer heterogeneities



Notes: Each pair of coefficients come from a single regression of an indicator for a positive review on an interaction between an indicator for whether the reviewer's prior review was positive and an indicator for a reviewer characteristic (e.g. female), controlling for paper fixed effects, reviewer fixed effects, dummies for number of reviews done prior to the current review, and a linear control for number of months since the reviewer's last review. Standard errors clustered at the reviewer level. Confidence intervals calculated at the 95% level.

Figure 2: Autocorrelation in reviewer recommendations - paper heterogeneities



Notes: Each pair of coefficients come from a single regression of an indicator for a positive review on an interaction between an indicator for whether the reviewer's prior review was positive and an indicator for a paper characteristic (e.g. all male authors), controlling for paper fixed effects, reviewer fixed effects, dummies for number of reviews done prior to the current review, and a linear control for number of months since the reviewer's last review. Standard errors clustered at the reviewer level. Confidence intervals calculated at the 95% level.