Decision-Making under the Gambler's Fallacy:

Evidence from Asylum Judges, Loan Officers, and Baseball Umpires *

Daniel Chen
ETH Zurich
Center for Law and Economics
chendan@ethz.ch

Tobias J. Moskowitz
University of Chicago
Booth School of Business
tobias.moskowitz@chicagobooth.edu

Kelly Shue
University of Chicago
Booth School of Business
kelly.shue@chicagobooth.edu

November 4, 2014

Abstract

Can misperceptions of what constitutes a fair process lead to unfair decisions? Previous research on the law of small numbers and the gambler's fallacy suggests that many people view sequential streaks of 0's or 1's as unlikely to occur even though such streaks often occur by chance. We hypothesize that the gambler's fallacy leads agents to engage in negatively autocorrelated decision-making. We document negatively autocorrelated decisions in three high-stakes contexts: refugee asylum courts, loan application review, and baseball umpire calls. This negative autocorrelation is stronger among more moderate and less experienced decision-makers, following longer streaks of decisions in one direction, and when agents face weaker incentives for accuracy. We show that the negative autocorrelation in decision-making is unlikely to be driven by potential alternative explanations such as sequential contrast effects, quotas, or preferences to treat two teams fairly.

Preliminary – Results Subject to Change

*Most recent version at: https://sites.google.com/site/kellyshue/research/. We thank Alex Bennett, Leland Bybee, Kaushik Vasudevan, Chattrin Laksanabunsong, Sarah Eichmeyer, and Luca Braghieri for excellent research assistance and Sue Long for helpful discussions about the asylum court data. We thank Stefano Dellavigna, Emir Kamenica, Josh Schwartzstein, and Dick Thaler for helpful comments. We thank seminar participants at ANU, Dartmouth, Rice, Rochester, SITE, U Chicago, U Indiana, U Oklahoma, U Rochester, U Washington, and UNSW for helpful comments.

1 Introduction

Research on the "law of small numbers" and the "gambler's fallacy" has well documented the tendency of people to overestimate the likelihood that a short sequence will resemble the general population (Tversky and Kahneman, 1971, 1974; Rabin, 2002; Rabin and Vayanos, 2010). For example, people may believe that a sequence of coin flips such as "01010" is more likely to occur than "00001" even though each sequence occurs with equal probability. Similarly, people may expect flips of a fair coin to generate high rates of alternation between 0's and 1's even though streaks of 0's or 1's often occur by chance. This misperception of random i.i.d. processes leads to errors in predictions: after observing one or more heads, the gambler feels that the fairness of the coin makes the next coin flip more likely to be tails.

Many of the existing empirical studies of the gambler's fallacy examines predictions in laboratory settings or betting errors in gambling markets (e.g. Benjamin, Moore, and Rabin, 2013; Ayton and Fischer, 2004; Croson and Sundali, 2005). In this paper, we show that the gambler's fallacy can bias high-stakes decision-making in real-world or field settings. Decision-makers such as judges, loan officers, umpires, HR interviewers, or auditors often make sequences of decisions under substantial uncertainty. We hypothesize that the gambler's fallacy leads agents to engage in negatively autocorrelated decision-making.

Our focus on decision-making differs from previous research on predictions because decisions can be made using both predictions about the quality of the next case as well as investigation of each case's merits. In fact, if the ordering of cases is random and decisions are made only based upon case merits, an agent's decision on the previous case should not predict the agent's decision on the next case, after controlling for base rates of affirmative decisions. However, a decision-maker who misperceives random processes may approach the next decision with a *prior* belief that the case is likely to be a 0 if she deemed the previous case to be a 1, and vice versa. This prior stems from the mistaken view that streaks of 0's and 1's are unlikely to occur by chance. Assuming that decisions made under uncertainty are at least partly influenced by the agent's priors, these priors will then lead to negatively autocorrelated decisions. Similarly, a decision-maker who fully understands random processes may still engage in negatively autocorrelated decision-making if she is being evaluated by others, such as promotion committees or voters, who suffer from the gambler's

fallacy.

We test our hypothesis in three high-stakes settings: refugee court asylum decisions in the US, a field experiment by Cole, Kanz, and Klapper (2013) in which experienced loan officers in India review real small-business loan applications in an experimentally controlled environment, and umpire calls of pitches in Major League Baseball games. In each setting, we show that the ordering of case quality is likely to be conditionally random. However, decisions are significantly negatively autocorrelated. We estimate that a significant percentage of decisions, more than 5 percent in some samples, are reversed due to the gambler's fallacy. We use the three settings to show that decision-making under the gambler's fallacy occurs in a wide variety of contexts and also because each setting offers unique benefits and limitations in terms of data analysis.

First, we test whether asylum judges are more likely to deny asylum after granting asylum to the previous applicant. The asylum courts setting offers administrative data on high frequency judicial decisions with very high stakes for the asylum applicants – judge decisions determine whether refugees seeking asylum will be deported from the US. The setting is also convenient because cases filed within each court (usually a city) are randomly assigned to judges within the court and judges decide on the queue of cases in a first-in-first-out fashion. By controlling for the recent approval rates of other judges in the same court, we are able to control for time-variation in court-level case quality to ensure that our findings are not generated spuriously by time variation in case quality. A limitation of the asylum court data is that we cannot discern whether any individual decision is correct given the case merits. However, we can estimate that up to two percent of decisions are reversed due to the gambler's fallacy. This effect is significantly stronger in certain subsamples: following a sequence of two decisions in the same direction, when judges have "moderate" grant rates close to 50% (calculated excluding the current decision), when the current and previous cases share similar characteristics or occur close in time (which is suggestive of coarse thinking as in Mullainathan et al., 2008). We also find that judge experience mitigates the negative autocorrelation.

Second, we test whether loan officers are more likely to deny a loan application after approving the previous application. The field experiment offers controlled conditions in which the order of loan files within each session is randomized by the experimenter. In addition, loan officers are randomly assigned to one of three incentive schemes, so we can test whether strong pay-for-performance reduces the bias in decision-making. The setting is also convenient in that we can observe true

loan quality, so we can discern loan officer mistakes. Finally, payoffs in the field experiment only depend on accuracy. Loan officers in the experiment are told that their decisions do not affect actual loan origination and they do not face quotas. Therefore, any negative autocorrelation in decisions is unlikely to be driven by concerns about external perceptions, quotas, or by the desire to treat loan applicants in a certain fashion. We find that up to 9 percent of decisions are reversed due to the gambler's fallacy in the flat incentive scheme among moderate decision-makers, although the effect is significantly smaller in the stronger incentive schemes and among less moderate decision-makers. Across all incentive schemes, the negative autocorrelation is stronger following a streak of two approval decisions in one direction. Finally, education, age, experience, and a longer period of time spent reviewing the current loan application reduces the negative autocorrelation in decisions.

Third, we test whether baseball umpires are more likely to call the current pitch a ball after calling the previous pitch a strike. An advantage of the baseball umpire data is that it includes precise measures of the trajectory and location of each pitch. Thus, while pitches may not be randomly ordered over time, we can control for each pitch's true location and measure whether mistakes in calls conditional on a pitch's true location is negatively predicted by the previous call. We find that umpires are 1.5 percentage points less likely to call a pitch a strike if the previous pitch was called a strike. This effect doubles when the current pitch is close to the edge of the strike zone (so it is a less obvious call) and following two previous calls in the same direction. We also show that any endogenous changes in pitch location over time are likely to be biases against our findings.

A potential interpretation issue that is specific to the baseball setting is that umpires may also have a preference to be equally nice or "fair" to two opposing teams. Such a desire is unlikely to drive behavior in the asylum judge and loan officers settings because the decision-makers review sequences of independent cases which are not part of teams. However, a preference to be equally nice to two opposing teams may lead to negative autocorrelation of umpire calls within a baseball inning. After calling a marginal or difficult-to-call pitch a strike, the umpire may choose to balance his calls by calling the next pitch a ball. We show that such preferences are unlikely to drive our estimates for baseball umpires. We find that the negative autocorrelation remains equally strong or stronger when the previous call was obvious (i.e. far from the strike zone boundary) and correct. In these cases, the umpire is less likely to feel guilt about making a particular call because the umpire probably could not have called the pitch any other way. Nevertheless, we find strong

negative autocorrelation following these obvious and/or correct calls, suggesting that a desire to undo marginal calls or mistakes is not the sole driver of our results.

Overall, we show that misperceptions of what constitutes a fair process and the desire to make correct calls can perversely lead to unfair decisions. Consistent with previous evidence showing that inexperience magnifies cognitive biases (Krosnick and Kinder, 1990; Chen and Berdejó, 2013), we find that education, experience, and strong incentives for accuracy can reduce biases in decisions caused by the gambler's fallacy. Our research also contributes the sizable psychology literature using vignette studies with small samples of judges that suggest unconscious heuristics (e.g., anchoring, status quo bias, availability) can play a large role in judicial decision-making (e.g. Guthrie et al., 2000).

We also consider potential alternative/complementary explanations. The first is sequential contrast effects (SCE), in which decision-makers perceive new information in contrast to what preceded it. Bhargava and Fisman (2012) find that subjects in a speed dating setting are more likely to reject the next candidate for a date if the previous candidate was very attractive. Under SCE, agents have a quality bar that moves following recent exposure to very high or low quality cases. We believe that SCE can be an important determinant of decision-making. However, we present a number of tests showing that SCE are unlikely to be a major driver of negatively autocorrelated decisions in our three empirical settings. In both the asylum court and loan approval settings, we find that an agent is not more likely to reject the current case if she approved a previous case that was very high in quality after conditioning on the previous decision. In the context of baseball pitches, there is an objective quality bar (the official strike zone) that should not move depending on the quality of the previous pitch.

A second potential alternative explanation is that agents face quotas for the number of affirmative decisions, which could also lead to negative autocorrelation in decisions. In all three of our empirical settings, agents do not face explicit quotas. For example, loan officers in the field experiment are only paid based upon accuracy and their decisions do not affect loan origination. However, one may be concerned about self-imposed quotas. For example, an asylum judge may wish to avoid granting asylum to too many applicants. We show that quotas are unlikely to explain our results by controlling for the fraction of the previous five or even two decisions that were decided in a certain direction. We find that, conditional on this fraction, extreme recency in the form of the previous

single decision still negatively predicts the next decision.

The next two potential explanations are closely related our gambler's fallacy hypothesis. Instead of attempting to rule them out, we present them as possible variants of our main hypothesis. The first is that the decision-maker is rational, but cares about the opinions of others, such as promotion committees or voters, who are fooled by randomness. These rational decision-makers will choose to make negatively-autocorrelated decisions in order to avoid the appearance of being too lenient or too harsh. We believe that concerns about external perceptions could be an important driver of decisions. However, they are unlikely to drive the results in the context of loan approval, which is an experimental setting where monetary payouts depend only on accuracy and the ordering of decisions and their associated accuracy is never reported to an outside party. The second related explanation is that agents may prefer to alternate being "mean" and "nice" over short time horizons. We cannot rule out this preference for mixing entirely. However, the desire to avoid being mean two times in a row, holding the overall fraction of negative decisions constant, could originate from the gambler's fallacy. A decision-maker who desires to be fair may over-infer that she is becoming too harsh and negative from a short sequence of "mean" decisions. Moreover, a preference to alternate mean and nice is again unlikely to drive behavior in the loan approval setting where loan officer decisions in the experiment know that they do not affect real loan origination (so there is no sense of being mean or nice to loan applicants).

Our paper builds upon the large body of work studying predictions under the gamblers fallacy.[1] Our focus on decisions highlights how the gambler's fallacy interacts with decision-making under uncertainty. Decisions differ from predictions in that decisions are based upon both prior beliefs (which can be biased by misperceptions of random processes) as well as information attained from

[1]Misperceptions of random processes can also lead to a related behavioral bias: the hot hand fallacy (Gilovich et al., 1985). In the hot hand fallacy, the agent is unsure of the mean of the population from which each observation is drawn or believes that the mean can be time varying. The agent holds an initial prior belief regarding the population mean. After observing a sequence of 1's (or 0's), the agent reasons that this sequence was unlikely to occur under the initial prior belief of the mean, and over-infers that the mean must be higher (lower) than initially expected, and therefore expects the streak to continue. For example, sports fans may be unsure of a basketball player's skill on a particular day. After observing a streak of shots, fans may overinfer that the basketball player's skill on that day is higher than initially expected, and expect him to make the next shot. Of course, players may indeed become hot; the hot hand fallacy refers to the overinference of skill from observations of streaks. The key differences between the hot hand and the gambler's fallacies are (1) the hot hand fallacy is more likely to occur when the agent is uncertain about the population mean, and (2) the hot hand fallacy only occurs after observing a longer streak of at least two draws while the gambler's fallacy can lead agents to expect reversals after a single draw (under the reasoning that another similar draw would lead to a streak, which is unlikely to occur). In unreported tests, we do not find strong evidence of the hot hand fallacy affecting decisions in our data.

reviewing the merits of each case. This implies that greater effort on the part of the decision-maker or better availability of information regarding the merits of the current case can reduce errors in decisions even if the decision-maker continues to suffer from the gambler's fallacy when forming predictions.

2 Model

To motivate why the gambler's fallacy may lead to negatively correlated decision-making, we present a simple extension of the Rabin (2002) model of coarse thinking. In the Rabin model, coarse thinkers believe that, within short sequences, black (1) and white (0) balls are drawn from an imaginary urn of finite size *without replacement*. Therefore, a draw of a black ball increases the odds of the next ball being white. As the size of the imaginary urn approaches infinity, the coarse thinker behaves like the rational thinker. We extend the model to decision-making by assuming that before assessing each case, agents hold a prior belief about the probability that the case will be a black ball. This prior belief is shaped by the same mechanics as the coarse thinker's beliefs in the Rabin model. However, the agent also receives a noisy signal about the quality of the current case, so the agent's ultimate decision is a weighted average of her prior belief and the noisy signal.

2.1 Model Setup

Suppose an agent makes 0/1 decisions for a randomly ordered series of cases. The true case quality is an i.i.d. sequence $\{y_t\}_{t=1}^{M}$ where $y_t = \{0, 1\}$, $P(y_t = 1) = \alpha \in (0, 1)$, and $y_t \perp y_{t-1}\ \forall t$.

The agent's prior about the current case is

$$P_t \equiv P\left(y_t = 1 \mid \{y_\tau\}_{\tau=1}^{t-1}\right).$$

For simplicity, we assume that the decision-maker believes the true case quality for all cases prior to t is equal to the decision made (e.g. if the agent decided the ball was black, she believes it is black).

The agent also observes a signal about current case quality $S_t \in \{0, 1\}$ which is accurate with probability μ and uninformative with probability $1 - \mu$. By Bayes Rule, the agent's belief after

observing S_t is

$$P\left(y_t = 1 \mid S_t, \{y_\tau\}_{\tau=1}^{t-1}\right) = \frac{[\mu S_t + (1-\mu)\alpha]\, P_t}{\alpha}.$$

The agent then imposes a threshold decision rule and makes a decision $D_t \in \{0, 1\}$ such that

$$D_t = 1\left\{\frac{[\mu S_t + (1-\mu)\alpha]\, P_t}{\alpha} \geq \overline{X}\right\}.$$

We then compare the prior beliefs and decisions of a rational agent to those of a coarse thinker. The rational agent understands that the y_t are i.i.d. Therefore, her priors are independent of history:

$$P_t^R = P\left(y_t = 1 \mid \{y_\tau\}_{\tau=1}^{t-1}\right) = P\left(y_t = 1\right) = \alpha.$$

By Bayes Rule, the rational agent's belief after observing S_t is

$$P\left(y_t = 1 \mid S_t = 1, \{y_\tau\}_{\tau=1}^{t-1}\right) = \mu S_t + (1-\mu)\,\alpha.$$

It is straightforward to see that the rational agent's decision on the current case should be uncorrelated with her decisions in previous cases, conditional on α.

In contrast, the coarse thinker believes that for rounds 1, 4, 7, ... cases are drawn from an urn containing N cases, αN of which are 1's (and the remainder are 0's). For rounds 2, 5, 8, ... cases are drawn from an urn containing $N-1$ cases, $\alpha N - y_{t-1}$ of which are 1's. Finally, for rounds 3, 6, 9, ... cases are drawn from an urn containing $N-2$ cases, $\alpha N - y_{t-1} - y_{t-2}$ of which are 1's. The degree of coarse-thinking is indexed by $N \in \mathbb{N}$ and we assume $N \geq 6$. As $N \to \infty$, the coarse-thinker behaves likes the rational thinker.

2.2 Model Predictions

The simple model generates the following testable predictions for coarse thinkers. For derivations, we refer the reader to Rabin (2002).

1. Decisions will be negatively autocorrelated.

2. "Moderate" decision-makers, defined as those with α close to 0.5, will make more unconditionally negatively autocorrelated decisions than extreme decision-makers, defined as those with

α close to 0 or 1.

3. The negative autocorrelation will be stronger following a streak of two or more decisions in the same direction.

4. The negative autocorrelation in decisions is stronger when the signal about the quality of the current case is less informative.

3 Empirical Framework

In this section, we describe the general empirical framework we will use across the three empirical contexts. In later sections when we describe each empirical setting in detail, we will discuss how the empirical specifications are customized to fit the unique needs of each setting.

3.1 Baseline

Our baseline specification tests whether the current decision is negatively correlated with the lagged decision:

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

Y_{it} represents binary decisions by decision-maker i ordered by t over time. β_1 measures the change in the probability of making an affirmative decision if the previous decision was an affirmative rather than a negative. If the ordering of cases is conditionally random, then $\beta_1 < 0$ is evidence in favor of the gambler's fallacy affecting decisions. While in some empirical settings, we cannot determine whether any particular decision was a mistake, we can use β_1 to estimate fraction of decisions that are reversed due to the gambler's fallacy: $2\beta_1 a\,(1-a)$, where a represents the base rate of affirmative decisions in the data.[2]

Even if the ordering of cases is random within each decision-maker, we face the problem that our estimate of β_1 may be biased upward when it is estimated using panel data with heterogeneity across

[2]Ignore the control variables for simplicity and consider the regression $Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \epsilon_{it}$. Let $a \equiv P\,(Y=1) = \beta_0/\,(1-\beta_1)$ be the base rate of affirmatives in the data. Suppose that absent the gambler's fallacy, the average approval rate would still equal a. If the previous decision was a negative, then the gambler's fallacy causes the current decision to be too likely to be an affirmative by the amount $(\beta_0 - a)$. If the previous decision was an affirmative, then the current decision is not likely enough to be an affirmative by the amount $(a - (\beta_0 + \beta_1))$. Therefore, the fraction of decisions that are reversed due to the gamblers fallacy is $(\beta_0 - a)\cdot P\,(Y_{i,t-1} = 0) + (a - (\beta_0 + \beta_1))\cdot P\,(Y_{i,t-1} = 1) = 2\beta_1 a\,(1-a)$.

decision-makers. The tendency of each decision-maker to be positive could be a fixed characteristic or slowly changing over time. This tendency to be positive can be thought of as a decision-maker specific α in the model which could also be slowly time varying. If we do not control for heterogeneity in α across decision-makers (and possibly within decision-makers over time), that would lead to upward bias for β_1 (a bias against us). This occurs because the previous decision and the current decision will both be positively correlated with the unobserved tendency to be positive.

We cannot control for α using decision-maker fixed effects. Within a finite panel, controlling for the mean within each panel leads to negative correlation between any two decisions by the same decision-maker. This biases toward $\beta_1 < 0$. Instead, we control for a moving average of the previous n decisions made by each decision-maker, not including the current decision. This tests whether the decision-maker reacts more to the most recent decision, controlling for the average grant rate among a recent set of decisions. In some tests, we also control for the decision-maker's average Y in settings other than the current setting (e.g. in other experimental sessions for the loan officers). Finally, we cluster standard errors by decision-maker or decision-maker×session as noted in later sections.

A second important reason we include control variables is that the sequence of cases considered in not necessarily randomly ordered within each decision-maker. To attribute $\beta_1 < 0$ to the gambler's fallacy, it must be true that the underlying quality of the sequence of cases considered, conditional on the set of controls, is not itself negatively autocorrelated. In the next sections, we discuss for each empirical setting why the sequences of cases are likely to be conditionally random. While we will present specific solutions in later sections, note that most types of non-random ordering in case quality correspond to persistent positive autocorrelation (e.g. trends in refugee quality) which would bias against findings of negative autocorrelation in decisions.

3.2 Streaks

We also test whether agents are more likely to reverse decisions following a streak of two or more decisions in the same direction. Specifically, we estimate

$$Y_{it} = \beta_0 + \beta_1 I(1,1) + \beta_2 I(0,1) + \beta_3 I(1,0) + Controls + \epsilon_{it}.$$

Here, $I(Y_{i,t-2}, Y_{i,t-1})$ is an indicator representing the two previous decisions. All β's measure behavior relative to the omitted group $I(0,0)$, in which the decision-maker has decided negatively two-in-a-row. A basic prediction of the gambler's fallacy model is that $\beta_1 < \beta_2 < 0$ and that $\beta_1 < \beta_3 < 0$. [3]All controls are as described in the baseline specification. However, we restrict our sample so that the current decision and as well as the two most recent decisions are consecutive.

4 Asylum Judges

4.1 Asylum Judges: Data Description and Institutional Context

The United States offers asylum to foreign nationals who can (1) prove that they have a well-founded fear of persecution in their own countries, and (2) that their race, religion, nationality, political opinions, or membership in a particular social group is one central reason for the threatened persecution. Decisions to grant or deny asylum have potentially very high stakes for the asylum applicants. An applicant for asylum reasonably fears imprisonment, torture, or death if forced to return to her home country. For a more detailed description of the asylum adjudication process in the US, we refer the interested reader to Ramji-Nogales et al. (2007).

We use administrative data on US refugee asylum cases considered in immigration courts from 1985 to 2013. Judges in immigration courts hear two types of cases: affirmative cases in which the applicant seeks asylum on her own initiative and defensive cases in which the applicant applies for asylum after being apprehended by the Department of Homeland Security (DHS). Defensive cases are referred directly to the immigration courts while affirmative cases pass a first round of review by asylum officers in the lower level Asylum Offices. The court proceeding at the immigration court level is adversarial and typically lasts several hours. Asylum seekers may be represented by an attorney at their own expense. A DHS attorney cross-examines the asylum applicant and argues before the judge that asylum is not warranted. Those that are denied asylum are ordered deported. Decisions to grant or deny asylum made by judges at the immigration court level are typically binding, although applicants may further appeal to the Board of Immigration Appeals.

Our baseline tests explore whether judges are less likely to grant asylum after granting asylum

[3]Under a strict interpretation of the Rabin (2002) coarse thinking model, we would also predict that $\beta_2 < \beta_3$, i.e. that extreme recency matters. In the model, this prediction requires assumptions regarding the probability that agents believe that the current case is the 1st, 2nd, or 3rd draw from the urn.

in the previous case. To attribute negative autocorrelation in decisions to the gambler's fallacy, we need to show that the underlying quality of the sequence of cases considered by each judge is not itself negatively autocorrelated. Several unique features of the immigration court process help us address this concern. Each immigration court covers a geographic region. Cases considered within each court are randomly assigned to the judges associated with the court (on average, there are eight judges per court). The judges then review the queue of cases following a "first-in-first-out" rule. In other words, judges do not exercise discretion in the order in which they review and decide on cases.[4]

Thus, any time variation in case quality (e.g. a surge in refugees from a hot conflict zone) should originate at the court-level. This variation in case quality is likely to be positively autocorrelated on a case-by-case level (later empirical tests support this claim). We also directly control for time-variation in court-level case quality using the recent approval rates of other judges in the same court.

Judges have a high degree of discretion in deciding case outcomes. They face no explicit or formally recommended quotas with respect to the grant rate for asylum. They are subject to the supervision of the Attorney General, but otherwise exercise independent judgment and discretion in considering and determining the cases before them. The lack of quotas and oversight is further evidenced by the wide disparities in grant rates among judges associated with the same immigration court (Ramji-Nogales et al., 2007).[5] Judges are attorneys appointed by the Attorney General as administrative judges. In our own data collection of immigration judge biographies, many judges previously worked as immigration lawyers or at the Immigration and Naturalization Service (INS) for some time before they were appointed. Judges typically serve until retirement. Their base salaries are set by a federal pay scale and locality pay is capped at Level III of the Executive Schedule. In 2014, that rate is $167,000. Based upon conversations with the President of the National Association of Immigration Judges, no bonuses are granted.

Our data comes from the Transactional Records Access Clearinghouse (TRAC). We exclude non-

[4]Exceptions to the first-in-first-out rule occur when applicants are heard multiple times, file applications on additional issues, get delays, or have closures made other than grant or deny (e.g. the applicant doesn't show up, withdraws, and an "other" category covering miscellaneous rare scenarios). We assume that these violations of first-in-first-out, which are likely driven by applicant behaviors, are uncorrelated with the judge's previous decision.

[5]For example, within the same four year time period in the court of New York, two judges granted asylum to fewer than 10% of the cases considered while three other judges granted asylum to over 80% of cases considered.

asylum related immigration decisions and focus on applications for asylum, asylum-withholding, or withholding-convention against torture. When an individual has multiple decisions on the same day on these three applications, we focus on one decision in the order listed above because the asylum decision applies to the asylum-withholding and withholding-convention against torture decisions and most individuals have all applications on the same day denied or granted. We merge this data with judicial biographies that we augmented. We exclude family members except the lead family member because in almost all cases, all family members are either granted or denied asylum together.

We also restrict our sample to decisions with known time ordering within day or across days and whose immediately prior decision by the judge is on the same day or previous day or over the weekend if it is a Monday decision. Finally, we restrict the sample to judges who have reviewed a minimum of 100 cases for a given court and courts with a minimum of 1000 cases in the data. Applying these exclusions restricts the sample to 150,357 decisions, covering 357 judges across 45 court houses.

Table 1
Asylum Judges: Summary Statistics

	Mean	Median	S.D.
Number of judges	357		
Number of courts	45		
Years since appointment	8.41	8	6.06
Daily caseload of judge	1.89	2	0.84
Family size	1.21	1	0.64
Grant indicator	0.29		
Non-extreme indicator	0.54		
Moderate indicator	0.25		
Lawyer indicator	0.939		
Defensive indicator	0.437		
Morning indicator	0.47		
Lunchtime indicator	0.38		
Afternoon indicator	0.15		

Table 1 summarizes our sample. Judges have long tenures, with a median of 8 years of experience. For data on tenure, we only have biographical data on 323 of the 357 judges, accounting for 142,699 decisions. The average caseload of a judge is 1.89 asylum cases per day. The average grant rate is 0.29. 94% of cases had a lawyer representing the applicant, and 44% were defensive cases, i.e. initiated by the government. The average family size is 1.21. 47% of hearings occurred in the

morning between 8 AM and 12 PM, 38% occurred during lunch time between 12 PM and 2 PM, and 15% occurred in the afternoon from 2 PM to 8 PM. We mark the clock time according to the time that a hearing session opened.

In later analysis, we show that the negative autocorrelation in judge decisions is driven by moderate judge observations, defined as those with grant rates for a given nationality-defensive category closer to 0.5 (calculated excluding the current observation). Note that the designation of moderate is not meant to imply that moderate judges are more reasonable or accurate in their decisions. Rather, the theory predicts that moderate decision makers should make more unconditionally negative decisions. Further, extreme decision makers who deny or grant asylum to all applicants within a category may believe that they hold very precise signals regarding whether a particular asylum applicant is worthy of asylum and therefore may rely less on prior beliefs that are biased by the gambler's fallacy. The non-extreme indicator tags decisions for which the average grant rate for the judge for that nationality-defensive category, calculated excluding the current observation, is between the 0.2 and 0.8. The moderate indicator tags decisions for which the average grant rate for the judge for that nationality-defensive category, calculated excluding the current observation, is between the 0.3 and 0.7.

4.2 Asylum Judges: Empirical Specification Details

Observations are at the judge x case order level. Y_{it} is an indicator for whether asylum is granted. Cases are ordered within day and across days. Our sample includes observations in which the lagged case was viewed in the same day or the previous workday (e.g. we include the observation if the current case is viewed on Monday and the lagged case was viewed on Friday). Observations in which there is a longer time gap between the current case and the lagged case are excluded from the sample. Multiple decisions on a single applicant are treated as one decision as they tend to be all "grants" or all "denies". Multiple family members are treated as one case for the same reason. Following Ramji-Nogales et al. (2007), we infer shared family status if cases share a hearing date, nationality, court, judge, decision, representation status, and case type (affirmative or defensive). Because our data contains some fields previously unavailable in the Ramji-Nogales et al. data, we also require family members to have the same lawyer identity code and to be heard during the same

or consecutive hearing start time.[6]

Control variables in the regressions include, unless otherwise noted, a set of dummies for the number of affirmative decisions over the past 5 decisions (excluding the current decision) of the judge. This controls for recent trends in grants, case quality, or judge mood. We also include a set of dummies for the number of affirmative decisions over the past 5 decisions across other judges (excluding the current judge) in the same court. This controls for recent trends in grants, case quality, or court mood. To control for longer term trends in judge- and court-specific grant rates, we control for the judge's average grant rate for the relevant nationality x defensive category, calculated excluding the current observation. We also control for the court's average grant rate for the relevant nationality x defensive category, calculated excluding the judge associated with the current observation. As noted previously, we don't include judge FE because that automatically induces negative correlation between Y_{it} and $Y_{i,t-1}$. Finally, we control for the characteristics of the current case: presence of lawyer representation indicator, family size, nationality x defensive fixed effects, and time of day fixed effects (morning / lunchtime / afternoon). The inclusion of time of day fixed effects is designed to control for other factors such as hunger or fatigue which may influence judicial decision-making (as shown in the setting of parole judges by Danziger et al., 2011).

4.3 Asylum Judges: Results

In Table 2, Column 1, we present results for the full sample of case decisions and find that judges are 0.5 percentage points less likely to grant asylum to the current applicant if the previous decision was an approve rather than a deny. In the remaining columns, we focus on cumulative subsamples in which the magnitude of the negative autocorrelation increases substantially. First, the asylum judges data cover a large number of judges who tend grant or deny asylum to almost all applicants from a

[6] A potential concern with inferring that two cases belong to the same family case using the criteria above is that family members must have, among the many other similarities, similar decision status. Therefore, sequential cases inferred to belong to different families will tend to have different decisions. This may lead to spurious measures of negative autocorrelation in decisions that is caused by error in the inference of families. We address this concern in two ways. First, we are much more conservative in assigning cases to families than Ramji-Nogales et al. (2007). In addition to their criteria, we also require family members to have the same identity for their lawyer and the same or consecutive hearing start time. This will lead to under-inference of families if some family members are seen during non-consecutive clock times or the data fails to record lawyer identity, both of which can occur in the data according to conversations with TRAC data representatives. Since family members tend to have the same decision, under-inference of families should lead to biases against our findings of negative autocorrelation in decisions. Second, we find evidence of significant negative autocorrelation when the current and previous case do not correspond to the same nationality. This type of negative autocorrelation cannot be generated by errors in the inference of families because family members will almost always have the same nationality.

certain nationalities. We do not claim in this paper that these more extreme judges are necessarily making incorrect decisions. However, these judges may exhibit less negative autocorrelation in their decisions because they believe they receive very precise signals regarding case quality, e.g. nationality X is never worthy of asylum. In Column 2, we restrict the sample to non-extreme judge observations (the average grant rate for the judge for that nationality-defensive category, calculated excluding the current observation, is between the 0.2 and 0.8). The extent of negative autocorrelation doubles to 1.1 percentage points. In unreported results, we find that autocorrelation in decisions among the omitted extreme judge observations sample is insignificant and close to zero. In Column 3, we restrict the sample to cases that follow another case on the same day. We find stronger negative autocorrelation within same-day cases and in unreported reports find near zero autocorrelation among sequential cases evaluated on different days. Column 4 restricts the sample to cases in which the current and previous case have the same defensive status, i.e. both defensive or both affirmative. The negative autocorrelation increases to 3.3 percentage points (among sequential cases with different defensive status, we again find close to zero autocorrelation). Finally, Column 5 tests whether decisions are more likely to be reversed following streaks of previous decisions. After a streak of two grants, judges are 5.5 percentage points less likely to grant asylum relative to decisions following a streak of two denials. Following a deny then grant decision, judges are 3.7 percentage points less likely to grant relative to decisions following a streak of two denials. Finally behavior following a grant then deny decision is insignificantly different from behavior following a streak of two denials.

Overall, we find evidence of significant negative autocorrelation in judge decisions. This negative autocorrelation is stronger among less extreme judges, when the current and previous case are less separated by time (occur in the same day), are more similar in terms of salient/defining characteristics (same defensive status), and following streaks of decisions in the same direction. These magnitudes are economically significant. For example, using the largest point estimate following a streak of two grant decisions: a 5.5 percentage point decline in the approval rate represents a 19% reduction in the probability of approval relative to the base rate of approval of 29 percent. Using the estimate in Column 4 within the sample of non-extreme, same-day, same defensive cases, the coefficient implies that approximately 1.5% of all decisions would have been reversed absent the gambler's fallacy.

Table 2
Asylum Judges: Baseline Results

This table tests whether the decision to grant asylum to the current applicant is related to the decision to grant asylum to the previous applicant. Observations are at the judge x case level. Observations are restricted to decisions that occurred within one day or weekend after the previous decision. Column 2 excludes extreme judge observations (the average grant rate for the judge for the nationality-defensive category of the current case, calculated excluding the current observation, is between 0.2 and 0.8). Column 3 further restricts the sample to decisions that follow another decision on the same day. Column 4 further restricts the sample to decisions in which the current and previous case have the same defensive status (both defensive or both affirmative). Column 5 tests how judges react to streaks in past decisions. *Lag grant-grant* is an indicator for whether the judge approved the two most recent asylum applicants. *Lag deny-grant* is an indicator for whether the judge granted the most recent applicant and denied the applicant before that. *Lag grant-deny* is an indicator for whether the judge denied the most recent applicant and granted the applicant before that. The omitted category is *Lag deny-deny.* All specifications include the following controls: indicator variables for the number of grants out of the judge's previous 5 decisions (excluding the current decision); indicator variables for the number of grants within the 5 most recent cases in the same court, excluding those of the judge corresponding to the current observation; the judge's average grant rate for the relevant nationality x defensive category (excluding the current observation); the court's average grant rate for the relevant nationality x defensive category (excluding the current judge); presence of lawyer representation indicator; family size; nationality x defensive fixed effects, and time of day fixed effects (morning / lunchtime / afternoon). Standard errors are clustered by judge. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Grant Asylum Dummy				
	(1)	(2)	(3)	(4)	(5)
Lag grant	-0.00544*	-0.0108***	-0.0155**	-0.0326***	
	(0.00308)	(0.00413)	(0.00631)	(0.00773)	
Lag grant - grant					-0.0549***
					(0.0148)
Lag deny - grant					-0.0367**
					(0.0171)
Lag grant - deny					-0.00804
					(0.0157)
Exclude extreme judges	No	Yes	Yes	Yes	Yes
Same day cases	No	No	Yes	Yes	Yes
Same defensive cases	No	No	No	Yes	Yes
N	150357	80733	36389	23990	10652
R^2	0.374	0.207	0.223	0.228	0.269

Table 3 explores additional heterogeneity. In this and subsequent tables, we restrict our analysis to the sample defined in Column 3 of Table 2, i.e. observations for which the current and previous case were decided by non-extreme judges on same day and had the same defensive status. Column 1 shows that the reduction in the probability of approval following a previous grant is 4.2 percentage points greater when the previous decision corresponds to an application with the same nationality

as the current applicant. While there is significant negative autocorrelation when sequential cases correspond to different applicant nationalities, the negative autocorrelation is three times larger when the two cases correspond to the same applicant nationality. This suggests that the gambler's fallacy may be tied to saliency and coarse thinking. Judges are more likely to engage in negatively autocorrelated decision-making when the previous case considered occurred close in time with the current case (as shown in the previous table) or was similar in terms of characteristics, i.e. nationality of the applicant or defensive status.

In the previous table, we found that non-extreme judges (those with grant rates, calculated excluding the current decision, between 20 and 80 percent) drove the bulk of the negative autocorrelation in decisions. In Column 2, we further show that moderate judges (those with grant rates, calculated excluding the current decision, between 30 and 70 percent) display stronger negative autocorrelation in decisions. Finally, Columns 3 and 4 show that judges who are inexperienced (less than 8 years of experience) are display stronger negative autocorrelation. Experience is associated with significantly reduced negative autocorrelation both cross sectionally (Column 3) and within judges over time (Column 4). [7]

Note that, because we measure decisions rather than predictions, reduced negative autocorrelation does not necessarily imply that some types of judge, e.g. experienced judges, are more sophisticated in terms of understanding random processes. Both experienced and inexperienced judges may suffer equally from the gambler's fallacy in terms of forming prior beliefs regarding the quality of the current case. However, experienced judges may draw, or believe they draw, more informative signals regarding the quality of the current case. If so, experienced judges will rely more on the current signal and less on their prior beliefs, leading to reduced negative autocorrelation in decisions.

[7] To identify the effect of experience within judges over time, we include judge fixed effects in Column 4. In general, we avoid inclusion of judge fixed effects because judge fixed effects bias the coefficient on *Lag grant* downward. However, the coefficient on *Lag grant x experienced judge* remains informative.

Table 3
Asylum Judges: Heterogeneity

Column 1 tests whether the gambler's fallacy is stronger when the previous decision concerned an applicant with the same nationality as the current applicant. Column 2 tests whether the gambler's fallacy is stronger among moderate judge observations (the average grant rate for the judge for the nationality-defensive category of the current case, calculated excluding the current observation, is between 0.3 and 0.7). Columns 3 and 4 test whether the gambler's fallacy declines with experience. Experienced in an indicator for whether the judge, at the time when the case was decided, had more than the median experience in the sample (8 years). Column 4 adds judge fixed effects, so the interaction term measures the within-judge effect of experience. All other variables and restrictions are as described in Table 2, Column 3. Standard errors are clustered by judge. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Grant Asylum Dummy			
	(1)	(2)	(3)	(4)
Lag grant	-0.0196**	0.00180	-0.0484***	-0.0553***
	(0.00801)	(0.00900)	(0.0115)	(0.0115)
Same nationality	0.0336***			
	(0.0108)			
Lag grant x same nationality	-0.0421***			
	(0.0126)			
Moderate judge		0.0326***		
		(0.0116)		
Lag grant x moderate judge		-0.0700***		
		(0.0136)		
Experienced judge			0.0138	0.0253*
			(0.0106)	(0.0140)
Lag grant x experienced judge			0.0327**	0.0456***
			(0.0152)	(0.0156)
Judge FE	No	No	No	Yes
N	23990	23990	22965	22965
R^2	0.229	0.229	0.229	0.247

Finally, we present evidence supporting the validity of our analysis. To attribute negative autocorrelation in decisions to the gambler's fallacy, we need to show that the underlying quality of the sequence of cases considered by each judge is not itself negatively autocorrelated. Within a court, the incoming queue of cases is randomly assigned to judges associated the court, and the judges review the queue of cases following a "first-in-first-out" rule. Therefore, time variation in case quality (e.g. a surge in refugees from a hot conflict zone) should originate at the court-level and is likely to be positively autocorrelated on a case-by-case level. We test this assumption in Table 4. In general, we find that case quality does not appear to be negatively autocorrelated in terms of observable proxies for quality, and if anything, is positively autocorrelated.

For each case, we create a predicted quality measure by regressing grant decisions on the following

case characteristics: whether the applicant had a lawyer, number of family members, whether the case warranted a written decision, and nationality x defensive status fixed effects. We estimate this regression using the entire sample of decisions, and create predicted grant status for each case using the estimated coefficients. Quality Measure 1 is this predicted grant status, normalized by the mean quality within the court in our sample. Quality Measure 2 is similar, except that we exclude all observations corresponding to the current judge from our prediction regression. This ensures that the current judge's history of decisions does not affect the creation of the predicted quality measure. We then regress these predicted quality measures on the lagged grant decision, using the same set of judge controls as in Table 2. We find no evidence of negative autocorrelation in case quality. Rather, following a previous grant decision, the next case is 0.3 percentage points more likely to be granted based upon observable measures. In the remaining columns of Table 4, we show that other case characteristics associated with higher grant rates (presence of a lawyer, average grant rate of the lawyer – calculated excluding the current case, and family size) do not decline following previous affirmative decisions. For these proxies of case quality, we estimate insignificant coefficients that are close to zero in magnitude. Overall, the underlying quality of cases appears to be slightly positively autocorrelated rather than negatively autocorrelated.

Table 4
Asylum Judges: Autocorrelation in Case Quality

This table tests whether lower quality cases tend to follow previous grant decisions. We create a predicted quality measure by estimating a first stage regression of grant decisions on case characteristics: whether the applicant had a lawyer, number of family members, whether the case warranted a written decision, and nationality x defensive status fixed effects. We estimate this regression using the entire sample of decisions and create predicted grant status for each case using the estimated coefficients. Quality Measure 1 is this predicted grant status, normalized by the mean predicted grant status within the court. Quality Measure 2 is similar, except the first stage regression is estimated excluding all observations corresponding to the current judge. Columns 1 and 2 regress these predicted quality measures on the lagged grant decision, using the same set of judge controls as in Table 2. Column 3 explores whether *Lag grant* is associated with higher probability of the next case having a lawyer. Column 4 explores whether *Lag grant* is associated with higher probability of the next case having a higher quality lawyer. Lawyer quality equals the average grant rate among cases represented by that lawyer, calculated excluding the current case. Cases without legal representation are excluded from this sample. Column 5 explores whether *Lag grant* is associated with the next case corresponding to a larger families (larger family size is positively associated with grant status). All other variables and restrictions are as described in Table 2. Standard errors are clustered by judge. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Quality Measure 1	Quality Measure 2	Lawyer Dummy	Lawyer Quality	Size of Family
	(1)	(2)	(3)	(4)	(5)
Lag grant	0.00273**	0.00307**	-0.0000772	-0.00117	-0.00927
	(0.00116)	(0.00134)	(0.00258)	(0.00293)	(0.0104)
N	23990	23980	23990	19737	23990
R^2	0.806	0.761	0.0858	0.451	0.159

5 Loan Officers

5.1 Loan Officers: Data Description and Institutional Context

We use field experiment data collected by Cole et al. (2013).[8] The original intent of the experiment was to explore how various incentive schemes affect the quality of loan officers' screening of loan applications. In the experiment, real loan officers were paid to screen actual loan applications. The framed field experiment was designed to closely match the underwriting process for unsecured small enterprise loans in India. Loan officers were recruited for the experiment from the active staff of several commercial banks. These loan officers had an average of 10 years of experience in the banking sector. In the field experiment, the loan officers screen real, previously processed loan

[8]For a detailed description of the data, we refer the interested reader to Cole et al. (2013). Our data sample consists of a subset of the data described in their paper. This data subsample was chosen by the original authors and given to us before any tests of the gambler's fallacy hypothesis were conducted. Therefore, differences between the subsample and full sample should not bias the analysis in favor of our findings.

applications. Each loan file contained all the information available to the bank at the time the loan was first evaluated.

Each loan officer participated in one or more evaluation sessions. In each session, the loan officer screened 6 randomly ordered loan files and decided whether to approve or reject the loan file. Because the loan files correspond to actual loans previously reviewed by banks in India, the files can be classified by the experimenter as performing or nonperforming. Performing loan files were approved and did not default in course of the actual life of the loan. Nonperforming loans were either rejected by the bank in the actual loan application process or were approved but defaulted in the actual life of the loan. Loan officers in the experiment were essentially paid based upon their ability to correctly classify the loans as performing (by approving them) or nonperforming (by rejecting them).

Participants in each session were randomly assigned to one of three incentive schemes which offered payouts of the form $[w_P, w_D, \overline{w}]$. w_P is the payout in rupees for approving a performing loan. w_D is the payout for approving a non-performing loan. $\overline{w}$ is the payout for rejecting a loan (regardless of actual loan performance). Beyond direct monetary compensation, participants may have also been motivated by reputational concerns. Loan officers were sent to the experiment by their home bank and the experiment was conducted at a loan officer training college. At the end of the experiment, loan officers received a completion certificate and a document summarizing their overall accuracy rate. The loan officers were told that this summary document would only report their overall accuracy without reporting the ordering of their specific decisions and associated accuracy. Thus, the loan officers may have been concerned that their home bank would evaluate these documents and therefore were motivated by factors other than direct monetary compensation. Importantly however, the ordering of decisions was never reported. Therefore, there was no incentive to negatively autocorrelate decisions to give the appearance of effort.

In the "flat" incentive scheme, payoffs take the form $[20, 20, 0]$, so loan officers had incentives to approve loans regardless of loan quality. The incentives in the "flat" scheme may at first seem surprisingly weak, but the authors of the original experiment used this incentive condition to mimic the relatively weak incentives faced by real loan officers in India. As shown in the next table, the overall approval rate within the flat incentive scheme is only 10 percentage points higher than the approval rates under the two other incentive schemes and loan officers were still more likely to

approve performing than nonperforming loans. This suggests that loan offers still chose to reject many loans and may have experienced some other intrinsic or reputational motivation to accurately screen loans.

In the "stronger" incentive scheme, payouts took the form $[20, 0, 10]$, so loan officers faced a monetary incentive to reject non-performing loans. In the "strongest" incentive scheme, payouts took the form $[50, -100, 0]$, so approval of non-performing loans was punished by deducting from a endowment given to the loan officers at the start of the experiment. The payouts across the incentive treatments were chosen to be approximately equal to 1.5 times the hourly wage of the median participant in the experiment.

The loan officers were also asked to assess the quality of each loan application on a 100 point scale. This quality score did not affect experimental payoffs, but Cole et al. (2013) show that the score is strongly predictive of loan approval and is correlated across different loan officers who reviewed the same loan file. The loan officers were informed of their incentive scheme. They were also made aware that their decision on the loans would affect their personal payout from the experiment but would not affect actual loan origination (because these were real loan applications that had already been evaluated in the past). Finally, the loan officers were told that the loan files were randomly ordered and that they were drawn from a large pool of loans of which approximately two-thirds were performing loans. Because the loan officers reviewed loans in an electronic system, they could not review the loans in any order other than the order presented. They faced no time limits or quotas.

Table 5
Loan Officers: Summary Statistics

	Full Sample		Flat Incentives		Strong Incentives		Strongest Incentives	
Loan officer x loan observations	9168		1332		6336		1470	
Loan officers	188		76		181		89	
Sessions (6 loans per session)	1528		222		1056		245	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Fraction loans approved	0.73		0.81		0.72		0.68	
Fraction moderate	0.34		0.25		0.36		0.36	
Loan rating (0-1)	0.71	0.16	0.74	0.16	0.70	0.16	0.73	0.15
Fraction grad school education	0.29		0.30		0.29		0.26	
Time viewed (minutes)	3.48	2.77	2.84	2.11	3.70	2.96	3.09	2.23
Age (years)	37.70	11.95	37.37	11.93	38.60	12.17	34.13	10.21
Experience in banking (years)	9.54	9.54	9.67	9.41	9.85	9.76	8.09	8.50

Table 5 presents summary statistics for our data sample. The data contains information on loan officer background characteristics such as age, education, and the time spent by the loan officer evaluating each loan file. Observations are at the loan officer x loan file level. We consider an observation to correspond to a moderate loan officer if the average approval rate of loans by the loan officer in other sessions (not including the current session) within the same incentive scheme is between 0.3 and 0.7. Again, our moderate classification does not imply that moderates are more accurate, but rather that the overall grant rate, excluding the current session, was closer to 0.5 than non-moderates.

5.2 Loan Officers: Empirical Specification Details

Observations are at the loan officer x loan level. Y_{it} is an indicator for whether the loan is approved. Loans are ordered within session. Our sample includes observations in which the lagged loan was viewed in the same session (so we exclude the first loan viewed in each session because we do not expect the gambler's fallacy to operate across sessions which are often separated by multiple days). In some specifications, we split the sample by incentive scheme type: flat, strong, or strongest.

As noted previously, we don't include loan officer fixed effects because that automatically induces negative correlation between Y_{it} and $Y_{i,t-1}$. Instead, to control for heterogeneity in mean approval rates at the loan officer x incentive scheme level, we control for the mean loan officer approval rate within each incentive treatment (calculated excluding the six observations corresponding to the current session). We also include an indicator for whether the loan officer has ever approved all six loans in another session within the same incentive treatment. Finally, we include an indicator for whether the current session is the only session attended by the loan officer within the incentive treatment (if so, the first two control variables cannot be calculated and are set to zero).

5.3 Loan Officers: Results

Table 6, Column 1, shows that loan officers are 8 percentage points less likely to approve the current loan if they approved the previous loan when facing flat incentives. This implies that 2.6 percent of decisions are reversed due to the gambler's fallacy. These effects become much more muted and insignificantly different from zero in the other incentive schemes when loan officers face stronger monetary incentives for accuracy. A test for equality of the coefficients indicate significantly

different effects across the three incentive schemes. In Column 2, we control for the true quality of the current loan file. Therefore, all reported coefficients represent mistakes on the the part of the loan officer. After including this control variable, we find qualitatively similar results.

In Columns 3 and 4, we repeat the analysis in the first two columns, but restrict the sample to loan officers with moderate approval rates (estimated using approval rates in other sessions excluding the current session). Comparing coefficients with those in the same row in Columns 1 and 2, we find that, within each incentive treatment, moderate decision-makers display much stronger negative autocorrelation in decisions. Under flat incentives, moderate decision makers are 23 percentage points less likely to approve the current loan if they approved the previous loan, implying that 9 percent of decisions are reversed due to the gambler's fallacy. Even within the stronger and strongest incentive treatments, loan officers are 5 percentage points less likely to approve the current loan if they approved the previous loan. One explanation for why the effect sizes are much larger in the moderate loan officers sample is that some loan officers may have decided to shirk in the experiment and approve almost all loans. Removing these loan officers from the sample leads to much larger effect sizes (note, we do not bias the results in favor of finding negative autocorrelation because we only use decisions in other loan sessions, excluding the current session, in the calculation of whether an observation corresponds to a moderate loan officer). Overall these tests suggest that loan officers, particularly moderate ones, exhibit significant negative autocorrelation in decisions which can be mitigated through the use of strong pay for performance.

Table 6
Loan Officers: Baseline Results

This table tests whether the decision to approve the current loan file is related to the decision to approve the previous loan file. Observations are at the loan officer x loan file level and exclude (as a dependent variable) the first loan file evaluated within each session. Columns 1 and 2 use the full sample while Columns 3 and 4 restrict the sample to moderate loan officers (an observation is considered moderate if the loan officer's average approval rate for loans, excluding the current session, is between 0.3 and 0.7 inclusive). Control variables include the loan officer's mean approval rate within each incentive treatment (calculated excluding the current session), an indicator for whether the loan officer has ever approved all six loans in another session within the same incentive treatment, and an indicator for whether the current session is the only session attended by the loan officer within the incentive treatment (if so, the first two control variables cannot be calculated and are set to zero). Indicator variables for flat incent, strong incent, and strongest incent are also included. Standard errors are clustered by loan officer x incentive treatment. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Approve Loan Dummy			
	(1)	(2)	(3)	(4)
Lag approve x flat incent	-0.0814**	-0.0712**	-0.225***	-0.228***
	(0.0322)	(0.0323)	(0.0646)	(0.0639)
Lag approve x stronger incent	-0.00674	-0.00215	-0.0525**	-0.0484**
	(0.0134)	(0.0134)	(0.0215)	(0.0214)
Lag approve x strongest incent	0.0102	0.0159	-0.0530	-0.0473
	(0.0298)	(0.0292)	(0.0468)	(0.0450)
p-value equality across incentives	0.0695	0.0963	0.0395	0.0278
Control for current loan quality	No	Yes	No	Yes
Sample	All	All	Moderates	Moderates
N	7640	7640	2615	2615
R^2	0.0257	0.0536	0.0247	0.0544

In the remaining analysis, we pool the sample across all three incentive treatments unless otherwise noted. Table 7 shows that loan officers with graduate school education and who spend more time reviewing the current loan file display significantly reduced negative autocorrelation in decisions.[9] Older and more experienced loan officers also display significantly reduced negative autocorrelation. These results are consistent with previous findings suggesting that education, experience, and effort can reduce behavioral biases. Note that, because we focus on decisions rather than predictions, our results do not necessarily imply that more educated, experienced, or conscientious loan officers suffer less from the gambler's fallacy. These loan officers may suffer strongly from the gambler's fallacy but draw, or believe they draw, more precise signals regarding current

[9]The sum of the coefficients on *Lag approve* and *Lag approve x grad approve* is positive, leading to the puzzling implication that loan officers with graduate school education engage in positively autocorrelated decision-making. However, our sample size is limited and the sum of the two coefficients is insignificantly different from zero.

loan quality, leading them to rely less on their priors regarding loan quality.

Table 7
Loan Officers: Heterogeneity

This table explores heterogeneity in the correlation between current and lagged decisions. Grad school is an indicator for whether the loan officer has a graduate school education. Time viewed is the number of minutes spent reviewing the current loan file. Age is the age of the loan officer in years. Experience is the loan officer's years of experience in the banking sector. All other variables are as described in Table 6. Standard errors are clustered by loan officer x incentive treatment. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Approve Loan Dummy			
	(1)	(2)	(3)	(4)
Lag approve	-0.0247*	-0.127***	-0.376***	-0.0555**
	(0.0135)	(0.0329)	(0.136)	(0.0250)
Grad school	-0.0213			
	(0.0214)			
Lag approve x grad school	0.0448*			
	(0.0245)			
Log(time viewed)		-0.0968***		
		(0.0202)		
Lag approve x log(time viewed)		0.0858***		
		(0.0230)		
Log(age)			-0.0603*	
			(0.0329)	
Lag approve x log(age)			0.101***	
			(0.0375)	
Log(experience)				-0.0133
				(0.00985)
Lag approve x log(experience)				0.0226*
				(0.0116)
Sample	All	All	All	All
N	7640	7640	7640	7640
R^2	0.0256	0.0281	0.0260	0.0256

Next, we test reactions to streaks of decisions. In Table 8, we find that after approving two applications in a row, loan officers are 7.5 percentage points less likely to approve the next application, relative to when the loan officer denied two applications in a row. After an approval, then rejection, the next decision is 6.9 percentage points more likely to be a rejection relative to when the officer made two rejections in a row. The effects are larger and more significant when restricted to moderate judges (Column 2). Note that "Lag reject - approve" has a less negative coefficient than "Lag approve - reject" even though a strict interpretation of the Rabin coarse thinking model would predict the opposite. The sample size is small, however, and the difference between these two coefficients is insignificant.

Table 8
Loan Officers: Reactions to Streaks

This table tests how loan officers react to streaks in past decisions. *Lag approve-approve* is an indicator for whether the loan officer approved the two most recent previous loans. *Lag approve-reject* is an indicator for whether the loan officer rejected the most recent previous loan and approved the loan before that. *Lag reject-approve* is an indicator for whether the loan officer approved the most recent previous loan and rejected the loan before that. The omitted category is *Lag reject-reject*, which is an indicator for whether the loan officer rejected the two most recent previous loans. The sample excludes observations corresponding to the first two loans reviewed within each session. All other variables are as described in Table 6. Standard errors are clustered by loan officer x incentive treatment. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Approve Loan Dummy	
	(1)	(2)
Lag approve - approve	-0.0751***	-0.165***
	(0.0216)	(0.0329)
Lag approve - reject	-0.0691***	-0.0955***
	(0.0236)	(0.0347)
Lag reject - approve	-0.0322	-0.0832**
	(0.0225)	(0.0332)
Sample	All	Moderates
N	6112	2092
R^2	0.0290	0.0322

We now discuss why our results are robust to a unique feature of the design of the original field experiment. Within each session, the order of the loans viewed by the loan officers on the computer screen was randomized. However, the original experimenters implemented a balanced session design. Each session consisted of four performing loans and two non-performing loans. If the loan officers had realized that sessions were balanced, a rational response would be to reject loans with greater probability after approving loans within the same session. There are two reasons why it is unlikely that loan officers would react to the balanced session design. First, they were not informed that sessions were balanced and were told that the loans were randomly selected from a large population of loans. Second, if loan officers had "figured out" that sessions were balanced, we would expect greater negative autocorrelation within the incentive treatments with stronger pay-for-performance. We find the reverse. In Columns 1 and 2 of Table 9, we reproduce the baseline results showing that the negative autocorrelation in decisions is strongest in the flat incentive scheme treatment. In Columns 3 and 4, we show that the negative autocorrelation in true loan quality is similar in magnitude across all three incentive treatments. Thus, if loan officers had realized that sessions were balanced, we would expect the opposite result, i.e. that the negative autocorrelation in decisions

would be equally or more strong under the stronger incentive schemes. In unreported analysis, we find similar results if we limit our sample to loan officers who participated in more than one incentive treatment. Finally, Table 5 shows that the approval rate of loans in the flat incentive treatment is 81% as compared to roughly 70% in the stronger and strongest incentive treatments. Again, if loan officers in the flat incentive scheme had realized that sessions were balanced with four performing loans and two non-performing loans, we would expect an approval rate closer to 66% instead of 81%. Overall, this evidence suggests that our baseline results are consistent with the gambler's fallacy and are unlikely to be generated by rational agents reacting to a balanced session design.

Table 9
Loan Officers: Robustness to Balanced Session Design

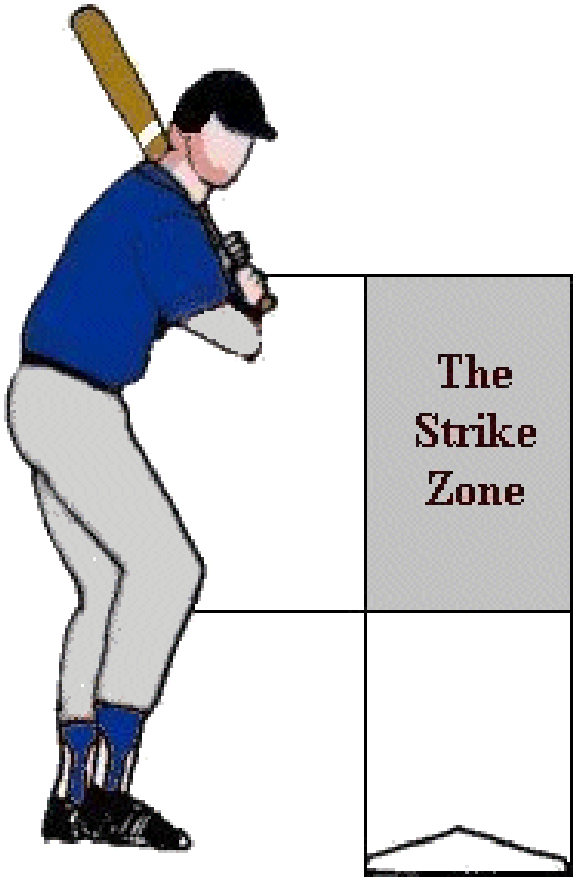
This table tests whether our results are robust to a balanced session design (each session consisted of 4 performing loans and 2 non-performing loans, randomly ordered). In Columns 1 and 2, we reproduce the results from Columns 1 and 3 of Table 6 showing that the negative autocorrelation in decisions is strongest under the flat incentive scheme. In Columns 3 and 4, we regress an indicator for the true quality of the current loan on the indicator for the true quality of the previous loan file. Indicator variables for flat incent, strong incent, and strongest incent are also included. All other variables are as described in Table 6. Standard errors are clustered by loan officer x incentive treatment. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Approve Loan Dummy		Performing Loan Dummy	
	(1)	(2)	(3)	(4)
Lag approve x flat incent	-0.0814**	-0.225***		
	(0.0322)	(0.0646)		
Lag approve x stronger incent	-0.00674	-0.0525**		
	(0.0134)	(0.0215)		
Lag approve x strongest incent	0.0102	-0.0530		
	(0.0298)	(0.0468)		
Lag perform x flat incent			-0.191***	-0.155***
			(0.0262)	(0.0529)
Lag perform x stronger incent			-0.131***	-0.142***
			(0.0123)	(0.0198)
Lag perform x strongest incent			-0.195***	-0.231***
			(0.0255)	(0.0407)
Sample	All	Moderates	All	Moderates
N	7640	2615	7640	2615
R^2	0.0257	0.0247	0.0235	0.0267

6 Baseball Umpires

6.1 Baseball Umpires: Data Description and Institutional Context

Figure 1
Baseball Umpires: The Strike Zone
According to Major League Baseball's "Official Baseball Rules" 2014 Edition, Rule 2.00, "The STRIKE ZONE is that area over home plate the upper limit of which is a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants, and the lower level is a line at the hollow beneath the kneecap. The Strike Zone shall be determined from the batter's stance as the batter is prepared to swing at a pitched ball."



In Major League Baseball (MLB), one important job of the umpire is to call a pitch as a either a strike or ball. If a batter does not swing, the umpire has to determine if the location of the ball as it passed home plate was within the strike zone shown in Figure 1. If the umpire decides the pitch is within the strike zone, he calls it a strike and otherwise calls it a ball. The boundaries of the strike zone are officially defined as in the caption for Figure 1, and are not subject to individual umpire interpretation. However, each umpire is expected to use his best judgment when determining the

location of the ball relative to the strike zone boundaries.

We use data on umpire calls of pitches from PITCHf/x, a system that tracks the trajectory and location of each pitch with respect to each batter's strike zone as the pitch crossed in front of home plate. The location measures are accurate to within a square centimeter. The Pitchf/x system was installed in 2006 in every Major League Baseball stadium. Our data covers approximately 3.5 million pitches over the 2008-2012 MLB seasons. We restrict our analysis to called pitches, i.e. pitches in which the batter does not swing (so the umpire must make a call), following a previous called pitch in the same inning. This sample restriction leaves us with approximately 1.5 million called pitches over 12,564 games by 127 different umpires. In some tests, we further restrict our sample to consecutive called pitches, i.e. the current called pitch and the previous called pitch were not interrupted by another pitch in which the umpire did not make a call (e.g. because the batter took a swing). Consecutive called pitches account for just under one million observations.

Table 10
Baseball Umpires: Summary Statistics

Number of called pitches following a previous called pitch	1536807
Number of called pitches following a consecutive previous called pitch	898741
Number of games	12564
Number of umpires	127
Fraction of pitches called as strike	0.3079
Fraction of pitches called correctly	0.8664
Fraction of pitches categorized as ambiguous	0.1686
Fraction of pitches categorized as obvious	0.3731
Fraction of ambiguous pitches called correctly	0.6006
Fraction of obvious pitches called correctly	0.9924

Table 10 summarizes our data sample. Approximately 30% of all called pitches are called as strikes (rather than balls). Umpires make the correct call approximately 85% of the time. We also categorize pitches by whether they were ambiguous (difficult to call) or obvious (easy to call). Ambiguous pitches fell ± 1.5 inches within the edge of the strike zone. For ambiguous pitches, 60% of umpire calls are called correctly. Obvious pitches fell within 3 inches around the center of the strike zone or 6 inches or more outside the edge of the strike zone. For obvious pitches, 99% of umpire calls are called correctly.

Our baseline tests explore whether umpires are less likely to call the current pitch a strike after calling the previous pitch a strike. To attribute negative autocorrelation in decisions to the gambler's fallacy, we need to assume that the underlying quality of the pitches (i.e. the location of the pitch relative to the strike zone), after conditioning on a set of controls, is not itself negatively autocorrelated. To address this potential concern, we include detailed controls for the characteristics of the current pitch: the pitch location relative to an absolute point on home plate (indicators for each 3x3 inch square), an indicator for whether the current pitch was within the strike zone, and the speed, acceleration, and spin in the x, y, and z directions of the pitch. For a complete detailed list of all control variables, please see the Appendix. These control variables address the concern that pitch characteristics are not randomly ordered. In addition, the fact that we control for whether the current pitch is actually within the true strike zone for each batter implies that any non-zero coefficients on other variables represent mistakes on the part of the umpire. Specifically, any coefficient on the lagged umpire decision will represent mistakes. In other words, we measure if the gambler's fallacy leads to mistaken calls in the *opposite* direction of the previous call, after controlling for the true location of the current pitch.

Of course, umpires may be biased in other ways besides the gambler's fallacy. For example, Parsons et al. (2011) show evidence of discrimination in calls: umpires are less likely to call strikes if the umpire and pitcher don't match in race and ethnicity. However, biases against teams or specific types of players should affect the base rate of called pitches within innings or against pitchers, and should not generate high-frequency negative autocorrelation in calls, which is the bias we focus on in this paper. More relevant for our findings, Moskowitz and Wertheim (2011) show that umpires may prefer to avoid making calls that strongly determine game outcomes. To focus on the gambler's fallacy as distinct from these types of biases, we control for indicator variables for every possible count combination[10] (# balls and strikes called so far for the batter), the leverage index (a measure developed by Tom Tango of how important a particular situation is in a baseball game depending

[10]We include indicator variables for every possible count combination to comprehensively control for other umpire biases that may depend on count. However, some count indicators completely determine the independent variable of interest, $Y_{i,t-1}$. For example, with a 0-1 count, 0 balls and 1 strike have been called so far, implying that $Y_{i,t-1}$ must equal 0. Observation following these determinative counts are used to more precisely estimate the other control variables. The coefficient on $Y_{i,t-1}$ is identified using observations in which $Y_{i,t-1}$ is not fully determined, e.g. calls following a 1-1 count in which the most recent call may have been a ball or strike. In unreported results, we find qualitatively similar coefficients on $Y_{i,t-1}$ if we do not control for count or control for count using continuous rather than indicator variables. In these specifications, the coefficient on $Y_{i,t-1}$ is identified using all observations.

on the inning, score, outs, and number of players on base), indicators for the score of the team at bat, indicators for the score of the team in the field, and an indicator for whether the batter belongs to the home team.

6.2 Baseball Umpires: Empirical Specification Details

The sample includes all called pitches except for the first in each game or inning. Y_{it} is an indicator for whether the current pitch is called a strike. $Y_{i,t-1}$ is an indicator for whether the previous pitch was called a strike. Control variables are as described in the previous section and detailed in the Appendix.

In this setting, we are particularly concerned that the "quality", i.e. location, of the pitch will also react to the umpire's previous call. We estimate a version of the analysis where the dependent variable is replaced with an indicator for whether the pitch was a true strike (within the strike zone). We also estimate a version of the analysis where the dependent variable is replaced with the distance of the pitch from the center of the strike zone. We test whether these proxies for the true location of the pitch depends on whether the lagged pitch was called a strike.

6.3 Baseball Umpires: Results

Table 11 Column 1 shows that umpires are 0.9 percentage point less likely to call a pitch a strike if the most recent previously called pitch was called a strike. Column 2 shows that the negative autocorrelation is stronger following streaks. Umpires are 1.3 percentage points less likely to call a pitch a strike if the two most recent called pitches were also called strikes. Further, umpires are less likely to call the current pitch a strike if the most recent pitch was called a strike and the pitch before that was called a ball than if the ordering of the last two calls were reversed. In other words, extreme recency matters. All analysis in this and subsequent tables include detailed controls for the actual location, speed, and curvature of the pitch. In addition, because we control for an indicator for whether the current pitch actually fell within the strike zone, all reported non-zero coefficients reflect mistakes on the part of the umpires (if the umpire always made the correct call, all coefficients other than the coefficient on the indicator for whether the pitch fell within the strike zone should equal zero).

Table 11
Baseball Umpires: Baseline Results

This table tests whether the decision to call the current pitch a strike is related to the decision to call the previous pitch(es) a strike. Observations are at the umpire x pitch level and exclude (as a dependent variable) the first pitch within each game. Columns 1 and 2 use the sample of all called pitches while Columns 3 and 4 restrict the sample to consecutive called pitches that are not interrupted by a pitch in which the umpire did not make a call (e.g. because the batter swung at the ball). Note that the sample size falls further in Column 4 because we require that the current pitch, previous pitch, and previous pitch before those are all consecutive. Control variables include the pitch location (indicators for each 3x3 inch square), an indicator for whether the current pitch was within the strike zone, the speed, acceleration, and spin in the x, y, and z directions of the pitch, break angle characteristics, indicators for every possible count combination (# balls and strikes called so far for the batter), the leverage index, indictors for the score of the team at bat and indicators for the score of the team in the field, an indicator for whether the batter belongs to the home team. For a complete detailed list of control variables, please see the Appendix. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Strike	Full Sample		Consecutive Pitches	
	(1)	(2)	(3)	(4)
Lag strike	-0.00919***		-0.0146***	
	(0.000591)		(0.000972)	
Lag strike - strike		-0.0131***		-0.0212***
		(0.00104)		(0.00268)
Lag ball - strike		-0.00994***		-0.0189***
		(0.000718)		(0.00156)
Lag strike - ball		-0.00267***		-0.00689***
		(0.000646)		(0.00155)
Pitch location	Yes	Yes	Yes	Yes
Pitch trajectory	Yes	Yes	Yes	Yes
Game conditions	Yes	Yes	Yes	Yes
N	1536807	1331399	898741	428005
R^2	0.669	0.668	0.665	0.669

In Columns 3 and 4 of Table 11, we repeat the analysis but restrict the sample to pitches that were called consecutively (so both the current and most recent pitch received umpire calls of strike or ball). In this restricted sample, the umpire's recent previous calls may be more salient because they are not separated by uncalled pitches. We find that the magnitude of the negative autocorrelation increases substantially in this sample. Umpires are 2.1 percentage points less likely to call the current pitch a strike if the previous two pitches were called strikes. This represents a 6.8 percent decline relative to the base rate of strike calls. In unreported results, we test whether the differences in magnitudes between the full sample and the consecutive called pitches sample are significant and find that they are with p-values below 0.001. In all subsequent analysis, unless

otherwise noted, we restrict the sample to consecutive called pitches.

Table 12
Baseball Umpires: Endogenous Pitcher Response

This table tests whether the location of the pitch relative to the strike zone is related to the decision to call the previous pitch(es) as strike. The sample is restricted to consecutive called pitches. The specifications are similar to those in Table 11, except that the dependent variable is replaced with a measure of pitch location. Columns 1 and 2 use an indicator for whether the current pitch was within the strike zone as the dependent variable. Columns 3-6 use the distance of the pitch in inches from the center of the strike zone as the dependent variable. Columns 1-4 exclude the following location control variables: pitch location (indicators for each 3x3 inch square) and an indicator for whether the current pitch was within the strike zone. Columns 5 and 6 use the full set of control variables, including location indicator variables, as described in Table 11. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	True Strike		Distance from Center			
	(1)	(2)	(3)	(4)	(5)	(6)
Lag strike	0.0168***		-0.275***		-0.00385	
	(0.00149)		(0.0236)		(0.00573)	
Lag strike - strike		0.0121***		-0.156**		-0.00403
		(0.00415)		(0.0701)		(0.0168)
Lag ball - strike		0.0200***		-0.361***		0.00651
		(0.00243)		(0.0367)		(0.00875)
Lag strike - ball		0.00308		-0.131***		0.00707
		(0.00241)		(0.0359)		(0.00854)
Pitch location	No	No	No	No	Yes	Yes
Pitch trajectory	Yes	Yes	Yes	Yes	Yes	Yes
Game conditions	Yes	Yes	Yes	Yes	Yes	Yes
N	898741	428005	898741	428005	898741	428005
R^2	0.0798	0.0924	0.171	0.188	0.952	0.952

Table 12 shows that the negative autocorrelation in umpire calls is unlikely to be caused by changes in the actual location of the pitch. We repeat the previous analysis but use measures of the current pitch's true location as our dependent variable. To identify the effect of previous calls on the location of the current pitch, we exclude location controls in Columns 1 - 4. Columns 1 and 2 use an indicator for whether the current pitch was within the strike zone as the dependent variable. If pitchers are more likely to throw true balls after the previous pitch was called a strike, we should find negative coefficients on lagged strike calls. Instead we find significant positive coefficients. In Columns 3 and 4, we use the distance of the pitch in inches from the center of the strike zone as the dependent variable. If pitchers are more likely to throw true balls (more distant from the center of the strike zone) after the previous pitch was called a strike, we should find significant positive coefficients on lagged strike calls; we find the opposite. These results imply that, following

a previous call of strike, the next pitch is likely to be closer to the center of the strike zone and another strike. In other words, endogenous changes in pitch location as a response to previous calls should lead to positive rather than negative autocorrelation in umpire calls. Finally, in Columns 5 and 6, we continue to use distance to the center of the strike zone as our dependent variable but now include the same set of detailed pitch location controls as in our baseline specifications. This is a test that our location controls account for close to all variation in pitch location. All reported coefficients on lagged calls become small and insignificantly different from zero.

Table 13
Baseball Umpires: Ambiguous vs. Obvious Calls
This table tests how our results differ depending on whether the current pitch is ambiguous or obvious. The sample is restricted to consecutive called pitches. Columns 1 and 2 restrict the sample to observations in which the current pitch is ambiguous (the location of the pitch is within 1.5 inches of boundary of the strike zone). Columns 3 and 4 restrict the sample to observations in which the current pitch is obvious (the location of the pitch is within 3 inches of the center of the strike zone or 6 inches or more outside of the edge of the strike zone. All control variables are as described in Table 11. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Strike	Current Pitch Ambiguous		Current Pitch Obvious	
	(1)	(2)	(3)	(4)
Lag strike	-0.0347***		-0.00226***	
	(0.00378)		(0.000415)	
Lag strike - strike		-0.0479***		-0.00515***
		(0.0113)		(0.00101)
Lag ball - strike		-0.0324***		-0.00442***
		(0.00566)		(0.000773)
Lag strike - ball		-0.000838		-0.00283***
		(0.00563)		(0.000841)
Pitch location	Yes	Yes	Yes	Yes
Pitch trajectory	Yes	Yes	Yes	Yes
Game conditions	Yes	Yes	Yes	Yes
N	151501	73820	335318	153996
R^2	0.317	0.316	0.891	0.896

Table 13 shows that the negative autocorrelation in decisions is reduced when umpires receive more informative signals about the quality of the current pitch, as predicted by the model. Columns 1 and 2 restrict the analysis to observations in which the current pitch is ambiguous, i.e. its location is close to the boundary of the strike zone. These pitches may be difficult to call due to their marginal locations. Columns 3 and 4 restrict the analysis to observations in which the current pitch is likely to be obvious, i.e. is location is close to the center of the strike zone or far from the edge of the strike

zone. We find that the magnitude of coefficients are ten to fifteen times larger when the current pitch is ambiguous relative to when the current pitch is obvious. In unreported analysis, we find that this difference in magnitudes is highly significant with p-values below 0.001. This is consistent with the model's predictions that the coarse thinker's prior beliefs about case quality will have less impact on the decision when the signal about current case quality is more informative.

Table 14
Baseball Umpires: Heterogeneity

This table tests how our results differ depending on game conditions or umpire characteristics. The sample is restricted to consecutive called pitches. Leverage and umpire accuracy are represented as z-scores. Leverage is a measure developed by Tom Tango of how important a particular situation is in a baseball game depending on the inning, score, outs, and number of players. Umpire accuracy is the fraction of pitches correctly called by the umpire, calculated excluding observations corresponding to the current game. High and low attendance are indicator variables for whether game attendance is in the highest and lowest quintiles of attendance, respectively (the omitted category consists of the middle three quintiles). All control variables are as described in Table 11. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)
Lag strike	-0.0146***	-0.0146***	-0.0143***
	(0.000972)	(0.000972)	(0.00108)
Leverage	0.000330		
	(0.000390)		
Lag strike x leverage	-0.00140**		
	(0.000625)		
Umpire accuracy		-0.00406***	
		(0.000451)	
Lag strike x umpire accuracy		0.00353***	
		(0.000621)	
High attendance			0.00441***
			(0.00115)
Low attendance			-0.00330***
			(0.00117)
Lag strike x high attendance			-0.00270*
			(0.00157)
Lag strike x low attendance			0.00123
			(0.00164)
Pitch location	Yes	Yes	Yes
Pitch trajectory	Yes	Yes	Yes
Game conditions	Yes	Yes	Yes
N	898741	898154	894779
R^2	0.665	0.665	0.665

In Table 14, we explore heterogeneity with respect to game conditions and umpire characteristics. Column 1 shows that an increase in leverage (the importance of a particular game situation for determining the game outcome) leads to significantly stronger negative autocorrelation in decisions.

However, the magnitude of the effect is small: a one standard deviation increase in game leverage leads to less than an 10 percent increase in the extent of negative autocorrelation. Column 2 shows that umpires who are more accurate (calculated as the fraction of pitches correctly called by the umpire in other games excluding the current game) also are less susceptible to the gambler's fallacy. A one standard deviation increase in umpire accuracy reduces negative autocorrelation by 25 percent. Finally, Column 3 tests whether the magnitude of negative autocorrelation varies by game attendance. We divide game attendance into quintiles and compare the highest and lowest quintiles to the middle three quintiles (which represent the omitted category). We don't find any significant differences in behavior by game attendance except in the highest quintile, where the negative autocorrelation increases by 18 percent. However, this difference in behavior is only marginally significant.

Table 15
Baseball Umpires: Treating Teams "Fairly"

This table tests whether our results are driven by umpires reversing previous marginal or incorrect calls. Columns 1 and 2 use the sample of all consecutive called pitches. Column 3 restricts the sample to pitches following a consecutive called pitch that was either obvious or ambiguous. *Prev call correct* and *prev call incorrect* are indicator variables for whether the umpire's previous call of strike or ball was correct or incorrect as measured by PITCHf/x. *Prev call obvious* in an indicator variable for whether the location of the previous called pitch was within 3 inches of the center of the strike zone or 6 inches or more outside of the edge of the strike zone. *Prev call ambiguous* is an indicator variable for whether the location of the previous pitch was within 1.5 inches of boundary of the strike zone. *Prev call not ambiguous/obvious* is an indicator equal to one if the previous pitch was neither obvious nor ambiguous. Column 3 further divides previous ambiguous calls by whether they were called correctly. This is not done for previous obvious calls because almost all, 99.3%, of obvious calls are called correctly as compared to 60.3% of ambiguous calls. In all columns, the reported interactions fully segment the regression sample. For example, the coefficient on "lag strike x prev call correct" represents the autocorrelation conditional on the previous call being correct and the coefficient on "lag strike x prev call incorrect" represents the autocorrelation conditional on the previous call being incorrect. All control variables are as described in Table 11. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Strike	Full Sample		Following Ambiguous/Obvious
	(1)	(2)	(3)
Lag strike x prev call correct	-0.0177***		
	(0.00101)		
Lag strike x prev call incorrect	-0.00663***		
	(0.00130)		
Lag strike x prev call obvious		-0.0180***	-0.0175***
		(0.00189)	(0.00216)
Lag strike x prev call ambiguous		-0.0120***	
		(0.00123)	
Lag strike x prev call not ambiguous/obvious		-0.0150***	
		(0.00103)	
Lag strike x prev call ambiguous and correct			-0.0140***
			(0.00175)
Lag strike x prev call ambiguous and incorrect			-0.00824***
			(0.00188)
Pitch location	Yes	Yes	Yes
Pitch trajectory	Yes	Yes	Yes
Game conditions	Yes	Yes	Yes
N	898741	895733	476819
R^2	0.665	0.665	0.666

An important consideration that is specific to the baseball setting is that umpires may have a preference to be equally nice or "fair" to two opposing teams and a desire to undo a previous marginal or mistaken call. The desire to be fair to two opposing teams is unlikely to drive results in the asylum judges and loan officers settings because the decision-makers review a sequence of independent cases, and the cases are not part of any teams. A preference to be equally nice or fair to two opposing teams may, however, drive the negative autocorrelation of umpire calls within an

baseball inning. After calling a marginal pitch a strike, the umpire may choose to balance out his calls by calling the next pitch a ball. While we cannot completely rule out these types of situations, we show that preferences for fairness are unlikely to drive our estimates for baseball umpires.

Table 15 Column 1 shows that negative autocorrelation is stronger following correct calls than following incorrect calls. This is inconsistent with a fairness motive, because umpires concerned with fairness should be more likely to reverse incorrect previous calls. Column 2 shows that the negative autocorrelation remains equally strong or stronger when the previous call was obvious (i.e. far from the strike zone boundary). In these cases, the umpire is less likely to feel guilt about making a particular call because the umpire could not have called it any other way. Nevertheless, we find strong negative autocorrelation following these obvious calls, suggesting that a desire to undo marginal calls is not the sole driver of our results. Finally, in Column 3, we restrict the sample to called pitches following previous calls that were either obvious or ambiguous. We further divide previous ambiguous calls into those that were called correctly (60%) and those that were called incorrectly (40%). If fairness concerns drive the negative autocorrelation in calls, the negative autocorrelation should be strongest following previous ambiguous and incorrect calls. We find the opposite. The negative autocorrelation is strongest following obvious calls (of which 99% are called correctly) and also stronger following previous ambiguous calls that were called correctly. Overall, these results suggest that fairness concerns and a desire to be equally nice to two opposing teams are unlikely to explain our results for baseball umpires.

7 Discussion and Alternative Explanations

7.1 Sequential Contrast Effects

Sequential contrast effects (SCE) describes situations in which the decision-maker's criteria for quality while judging the current case is higher if the previous case was particularly high quality. For example, after reading a really great book, one's standard for judging the next book to be "good" on a 0/1 scale may be higher. Like the gambler's fallacy, SCE can lead to negative autocorrelation in decisions.

We believe that SCE can be an important determinant of decision-making. However, we present a number of tests showing that SCE are unlikely to be a major driver of negatively autocorrelated

decisions in our three empirical settings. First, SCE are unlikely to occur in the context of baseball umpires in which there is an well-defined quality bar: did the pitch fall inside or outside the regulated strike zone?

Second, we can estimate:

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

This is the same as our previous specification except that we also introduce a continuous measure of quality for the previous case. If SCE drives the our findings, then we expect to find that $\beta_2 < 0$. Holding constant the previous discrete decision $Y_{i,t-1}$, decision-makers should be more likely to reject the current case if the previous case was of high quality, as measured continuously using $Quality_{i,t-1}$.[11]

Table 16 shows that sequential contrast effects are unlikely to drive our results in the case of asylum judges. For each case, we use the continuous predicted quality measure as described in Table 4. We then include the lagged case's predicted quality measure as a control variable. When we control for both the continuous quality of the lagged case and actual lag decision, the current decision is negatively correlated with the previous decision, but weakly positively correlated with the continuous proxy for the previous case's quality. We can reject that $\beta_2 < 0$ with p-values below 0.1. This is inconsistent with sequential contrast effects driving our results.

[11] Under a simple model of the gambler's fallacy in decision-making, agent react negatively to the previous binary decision. In a more nuanced model of the gambler's fallacy, such as that proposed in Rabin and Vayanos (2010), agents may react more negatively to previous affirmative decisions if they are more certain that the previous case was a true "1" because it was high in quality. Such a model would also predict that $\beta_2 < 0$. Empirically we find that β_2 is close to zero or slightly positive, contrary to the predictions of both the SCE model and this more nuanced model the gambler's fallacy.

Table 16
Asylum Judges: Sequential Contrast Effects?

This table tests whether the negative correlation between current asylum grant and lagged asylum grant could be caused by sequential contrast effects. "Lag Case Quality" is a continuous measure of the quality of the most recently reviewed asylum case as defined in Table 4 while Lag Grant is a binary measure of whether the previous asylum was granted. Conditional on the binary measure of whether the previous asylum was granted, sequential contrast effects predict that the judge should be less likely to grant asylum to the current applicant if the previous case was of higher quality, measured continuously. In other words, sequential contrast effects predicts that the coefficient on "Lag Grant Quality" should be negative. Standard errors are clustered by judge. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Grant Asylum Dummy	
	(1)	(2)
Lag grant	-0.0356***	-0.0352***
	(0.00788)	(0.00785)
Lag case quality	0.0394*	0.0285
	(0.0219)	(0.0197)
p-value lag case quality < 0	0.0367	0.0751
Quality Measure	1	2
N	23981	23973
R^2	0.228	0.228

Table 17 presents a similar test in the context of loan officers. As part of the field experiment, loan officers reported their assessment of loan quality on a 0 to 100 point scale. These scores did not directly affect experiment payoffs, but Cole et al. (2013) shows these scores are correlated with loan approval decisions and are also consistent across different loan officers who reviewed the same loan file. This evidence suggests that these scores reflect loan officers' perceptions of loan quality. We again find evidence contrary to the predictions of a sequential contrast model – we can reject that $\beta_2 < 0$ with p-values around 0.1.

Table 17
Loan Officers: Sequential Contrast Effects?

This table tests whether the negative correlation between current loan approval and lagged loan approval could be caused by sequential contrast effects. "Lag Loan Quality Rating" is a continuous measure of the quality of the most recently reviewed loan file while Lagged Approve is a binary measure of whether the previous loan was approved. Conditional on the binary measure of whether the previous loan was approved, sequential contrast effects predict that the loan officer should be less likely to approve the current loan if the previous loan was of higher quality, measured continuously. In other words, sequential contrast effects predicts that the coefficient on "Lag Loan Quality Rating" should be negative. The loan quality measure is rescaled to vary from 0 to 1. All other variables are as described in Table 6. Standard errors are clustered by loan officer x incentive treatment. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Approve Loan Dummy	
	(1)	(2)
Lag approve	-0.0218*	-0.0854***
	(0.0126)	(0.0241)
Lag loan quality rating	0.0642*	0.124
	(0.0380)	(0.101)
p-value lag loan quality rating < 0	0.0458	0.109
Sample	All	Moderates
N	7640	2615
R^2	0.0254	0.0230

7.2 Quotas

A second potential explanation for negatively autocorrelated decision-making is that agents face quotas for the total number of positive decisions. In all three of our empirical settings, agents do not face explicit quotas. For example, loan officers are paid based upon accuracy and are explicitly told that they do not face quotas. However, one may be concerned that decision makers face self-imposed quotas. For example, an asylum judge may wish to avoid granting asylum to too many applicants. We show that self-imposed quotas are unlikely to explain our results by controlling for the fraction of the previous two or five decisions that were called in a certain direction. Controlling for the fraction of the previous 5 decisions decided in the affirmative, extreme recency in the form of the previous single decision still negatively predicts the next decision. Similarly, we control for the fraction of the previous two decisions granted and test whether the previous single decision still matters. We continue to find that the previous single decision negatively predicts the next decisions, with the exception of the loan officers field experiment in which the coefficients on *Lag grant-deny* and *Lag deny-grant* do not significantly differ from one-another, potentially due to the

smaller sample size. In general, agents negatively react to extreme recency holding the fraction of previous 5 or 2 decisions granted constant. This behavior is consistent with models of gambler's fallacy. It is also largely inconsistent with self-imposed quotas, unless the decision-maker also has limited memory and cannot remember the previous two decisions.

7.3 External Perceptions and Preferences for Alternation

We now discuss two potential explanations for negatively-autocorrelated decisions that are closely related to our gambler's fallacy hypothesis. Instead of attempting to rule them out, we present them as possible variants of our main hypothesis. The first is that the decision-maker fully understands random processes, but cares about the opinions of others, such as promotion committees or voters, who are fooled by randomness. These rational decision-makers will choose to make negatively correlated decisions in order to avoid the appearance of being too lenient or too harsh. We believe concerns about external perceptions could be an important driver of decisions. However, they are unlikely to drive the results in the context of loan approval, which is an experimental setting where payouts depend only on accuracy and the ordering of decisions and their associated accuracy is never reported to participants or their home banks. The second related explanation is that agents may prefer to alternate being "mean" and "nice" over short time horizons. We cannot rule out this preference for mixing entirely. However, the desire to avoid being mean two times in a row, holding the overall fraction of negative decisions constant, could actually originate from the gambler's fallacy. A decision-maker who desires to be fair may over-infer that she is becoming too harsh and negative from a short sequence of "mean" decisions. Moreover, a preference to alternate mean and nice is again unlikely to drive behavior in the loan approval setting where loan officer decisions in the experiment do not affect real loan origination (so there is no sense of being mean or nice).

8 Conclusion

We show that misperceptions of what constitutes a fair process can perversely lead to unfair decisions. Previous research on the law of small numbers and the gambler's fallacy suggests that many people view sequential streaks of 0's or 1's as unlikely to occur even though such streaks often occur by chance. We hypothesize that the gambler's fallacy leads agents to engage in nega-

tively autocorrelated decision-making. We document negative autocorrelation by decision-makers in three high-stakes contexts: refugee asylum courts, loan application review, and baseball umpire calls. This negative autocorrelation is stronger among more moderate and less experienced decision-makers, following longer streaks of decisions in one direction, and when agents face weaker incentives for accuracy. Finally, we show that the negative autocorrelation in decision-making is unlikely to be driven by potential alternative explanations such as sequential contrast effects, quotas, or preferences to treat two teams fairly.

References

Ayton, Peter, and Ilan Fischer, 2004, The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness?, *Memory & cognition* 32, 1369–1378.

Benjamin, Daniel, Don Moore, and Matthew Rabin, 2013, Misconceptions of chance: Evidence from an integrated experiment., *Working Paper* .

Bhargava, Saurabh, and Ray Fisman, 2012, Contrast effects in sequential decisions: Evidence from speed dating, *Review of Economics and Statistics* .

Chen, Daniel L., and Carlos Berdejó, 2013, Priming ideology? electoral cycles without electoral incentives among elite u.s. judges, Technical report, ETH Zurich, Mimeo.

Cole, Shawn, Martin Kanz, and Leora Klapper, 2013, Incentivizing calculated risk-taking: Evidence from an experiment with commercial bank loan officeres, *forthcoming Journal of Finance* .

Croson, Rachel, and James Sundali, 2005, The gambler's fallacy and the hot hand: Empirical data from casinos, *Journal of Risk and Uncertainty* 30, 195–209.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso, 2011, Extraneous factors in judicial decisions, *Proceedings of the National Academy of Sciences* 108, 6889–6892.

Gilovich, Thomas, Robert Vallone, and Amos Tversky, 1985, The hot hand in basketball: On the misperception of random sequences, *Cognitive Psychology* 17, 295–314.

Guthrie, Chris, Jeffrey J. Rachlinski, and Andrew J. Wistrich, 2000, Inside the judidical mind, *Cornell Law Review* 86, 777–830.

Krosnick, Jon A., and Donald R. Kinder, 1990, Altering the foundations of support for the president through priming, *The American Political Science Review* 84, 497–512.

Moskowitz, Tobias, and L. Jon Wertheim, 2011, *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won* (Crown Publishing Group).

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer, 2008, Coarse thinking and persuasion, *The Quarterly Journal of Economics* 123, 577–619.

Parsons, Christopher, Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh, 2011, Strike three: Discrimination, incentives, and evaluation., *American Economic Review* 101, 1410–35.

Rabin, Matthew, 2002, Inference by believers in the law of small numbers, *The Quarterly Journal of Economics* 117, 775–816.

Rabin, Matthew, and Dmitri Vayanos, 2010, The gambler's and hot-hand fallacies: Theory and applications., *Review of Economic Studies* 77, 730–778.

Ramji-Nogales, Jaya, Andrew I Schoenholtz, and Philip G Schrag, 2007, Refugee roulette: Disparities in asylum adjudication, *Stanford Law Review* 295–411.

Tversky, Amos, and Daniel Kahneman, 1971, Belief in the law of small numbers., *Psychological bulletin* 76, 105.

Tversky, Amos, and Daniel Kahneman, 1974, Judgment under uncertainty: Heuristics and biases, *Science* 185, 1124–1131.

Appendix

The empirical tests for baseball umpire decisions include the following control variables unless otherwise noted. All controls are introduced as linear continuous variables unless otherwise specified below.

1. Indicator variables for each 3×3 inch square for the (x, y) location of the pitch as it passed home plate, with $(0, 0)$ being lowest left box from perspective of umpire
2. Indicator for whether the batter belongs to the home team
3. Indicator for each possible pitch count combination (number of balls and strikes prior to current pitch)
4. Acceleration of the pitch, in feet per second per second, in the x-, y-, and z- direction measured at the initial release point (three continuous variables)
5. Break angle: The angle, in degrees, from vertical to the straight line path from the release point to where the pitch crossed the front of home plate, as seen from the catcher's/umpire's perspective
6. Break length: The measurement of the greatest distance, in inches, between the trajectory of the pitch at any point between the release point and the front of home plate, and the straight line path from the release point and the front of home plate
7. The distance in feet from home plate to the point in the pitch trajectory where the pitch achieved its greatest deviation from the straight line path between the release point and the front of home plate
8. End speed: The pitch speed in feet per second measured as it crossed the front of home plate
9. The horizontal movement, in inches, of the pitch between the release point and home plate, as compared to a theoretical pitch thrown at the same speed with no spin-induced
10. The vertical movement, in inches, of the pitch between the release point and home plate, as compared to a theoretical pitch thrown at the same speed with no spin-induced movement

11. The left/right distance, in feet, of the pitch from the middle of the plate as it crossed home plate (The PITCHf/x coordinate system is oriented to the catcher's/umpire's perspective, with distances to the right being positive and to the left being negative)

12. The height of the pitch in feet as it crossed the front of home plate

13. The direction, in degrees, of the ball's spin. A value of 0 indicates a pitch with no spin. A value of 180 indicates the pitch was spinning from the bottom

14. Spin rate: The angular velocity of the pitch in revolutions per minute

15. The velocity of the pitch, in feet per second, in the x, y, and z dimensions, measured at the initial point (three continuous variables)

16. The left/right distance, in feet, of the pitch, measured at the initial point

17. The height, in feet, of the pitch, measured at the initial point

18. Proportion of previous pitches to the batter during the given game that were either in the dirt or were a hit by pitch

19. Proportion of previous pitches to the batter during the given game that were put into play

20. Proportion of previous pitches to the batter during the game that were described as either swinging strike, missed bunt or classified as strike

21. Proportion of previous pitches to the batter during the game that were described as either intentional ball, pitchout, automatic ball, or automatic strike

22. Proportion of previous pitches to the batter during the game described as foul tip, foul, foul bunt, foul (runner going) or foul pitchout

23. Proportion of previous pitches to the batter during the game described as “ball".

24. Proportion of previous pitches to the batter during the game described as “called strike"

25. Indicator variable for whether the pitch should have been called a strike based on the objective definition of the strike zone

26. A measure developed by Tom Tango of how important a particular situation is in a baseball game depending on the inning, score, outs, and number of players on base

27. Indicator variables for each possible score of the team at bat

28. Indicator variables for each possible score of the team in the field