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

August 31, 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 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. We estimate that the gambler's fallacy reverses up to 5% of all decisions. 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.

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). 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. 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. We hypothesize that the gambler's fallacy leads agents to engage in negatively autocorrelated decision-making. Decision-makers such as judges, loan officers, umpires, HR interviewers, or auditors often make sequences of decisions under substantial uncertainty. If the ordering of cases is random, an agent's decision on the previous case should not predict the agent's decision on the next case if decisions are made based upon case merits.[1] 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 et al. (2013) in which experienced loan officers in India review real small-business loan applications in an experimentally controlled environment, and umpire calls of pitches

[1] This is assuming that we control for the base rate of making an affirmative decision.

in Major League Baseball games. In each setting, we show that the ordering of cases is likely to be conditionally random. However, decisions are significantly negatively autocorrelated. We estimate that up to 5 percent of decisions 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 in that cases filed within each court (usually a city) are randomly assigned to judges within the court and judges must 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 are busy or have "moderate" grant rates close to 50%, when the current and previous cases share similar characteristics (which is suggestive of coarse thinking as in Mullainathan et al., 2008), and when the current and previous decisions occur close in time. 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 8 percent of decisions are reversed due to the gambler's fallacy in the flat incentive scheme, although the effect becomes insignificant in the stronger incentive schemes. Across all incentive schemes, the negative autocorrelation is stronger among more moderate loan officers (those who approve close to 50% of loans in other sessions excluding the current session) and 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 the gambler's fallacy leads to mistakes in umpire calls. We find that up to 1.5% of all calls are mistakes caused by the gambler's fallacy. This effect increases to 3.5% when the current pitch is close to the edge of the strike zone (so it is a less obvious call) and to 5% 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 concern 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 an baseball inning. After calling a marginal or difficult-to-call pitch a strike, the umpire may choose to balance out 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 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 probably 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.

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. Like the gambler's fallacy, sequential contrast effects can lead to negative autocorrelation in binary 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. In both the asylum court and loan approval settings, we find that an agent is more likely to reject the current case if she approved a previous case that was moderate in quality than if she approved a previous case that was very high in quality. This result is the opposite of that predicted by SCE. In the context of baseball pitches, we find, as discussed above, no significant difference in the probability of calling the current pitch a strike if the previous pitch was very obviously a strike (high quality) or less obviously a strike (low quality).

A second potential alternative explanation is that agents face a quota for the number of affirmative decisions they can grant, 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 5, 10, or set of calls within a baseball inning that were called in a certain direction. We find that, conditional on these controls, extreme recency in the form of the previous single decision still negatively predicts the next decision.

The next two potential explanations for negatively-autocorrelated decisions 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 payouts depend only on accuracy. The second related explanation is that agents may prefer to alternate being "mean" and "nice" over short time horizons. This can be viewed as a preference for sequences that follow the law of small numbers rather than a belief in the law of small numbers. We cannot rule out this preference for mixing entirely. However, it 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) and subjects are paid purely for accuracy.

Our paper builds upon the large body of work studying predictions under the gamblers fallacy.[2] 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 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.

[2]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 population 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 population mean, and over-infers that the population 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 significant evidence of the hot hand fallacy affecting decisions in our data.

2 Model

To motivate why the gambler's fallacy may lead to negatively correlated decision-making, we present a very 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 Rabin coarse thinking model to a model of 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 receive 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

More formally, 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

$$D_t = 1\left\{\frac{[\mu S_t + (1-\mu)\alpha]\, P_t}{\alpha} \geq 1/2\right\}.$$

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 four 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 specifications 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. If the ordering of cases is conditionally random, then $\beta_1 < 0$ is evidence in favor of the gambler's fallacy affecting decisions. We can also interpret β_1 to represent the fraction of decisions that are distorted due to the gambler's fallacy. In other words, β_1 represents the fraction of decisions that would have gone the other way if not for the gambler's fallacy.[3]

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

[3] Ignore the control variables for simplicity, and consider the regression $Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \epsilon_{it}$. Applying expectations to both sides, $P(Y_{it} = 1) = \beta_0 + \beta_1 P(Y_{i,t-1} = 1)$. Rearranging terms yields $P(Y_{it} = 0) = 1 - \beta_0 - \beta_1 + \beta_1 P(Y_{i,t-1} = 0)$. This implies that the fraction of all decisions that are distorted due to the gambler's fallacy is β_1.

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 instead 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 slow-moving 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 $\beta_1 < \beta_3 < 0$. [4]All controls are as described in the baseline specification. However, we restrict our

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

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 at last 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 1987 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. A DHS attorney cross-examines the asylum applicant and argues before the judge that asylum is not warranted. Asylum seekers may be represented by an attorney at their own expense. 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. Those that are denied asylum are ordered deported.

Our baseline tests explore whether judges are less likely to grant asylum after granting asylum 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.[5]

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. We also directly control for time-variation in court-level case quality using the recent approval rates of other judges in the same court.

Our data comes from the Transactional Records Access Clearinghouse (TRAC). The data contains the exact time when a decision was made, the identity of the judge, and litigant characteristics. We know when the asylum case was assigned, whether the hearing was an individual hearing or whether multiple individuals were scheduled in the same session, how many cases were scheduled for sessions during a day for each judge, whether this was an in person hearing or by audio or video, whether it was a written or oral order, whether there are other related applications for relief filed by the individual and the judge's ruling on each, the ethnicity of the applicant, whether the case was filed in the defensive or affirmative, and the identity of the judge including a limited set of judge demographic characteristics.

We merge the decision and hearing datasets. We exclude non-asylum decisions: that is, we focus on applications for asylum, asylum-withholding, or withholding-convention against torture. When an individual had multiple decisions on the same day on these three applications, we focus on one decision in the order listed above, as asylum decisions would be the most deterministic in terms of deportation 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.

Finally, we restrict our sample to decisions 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. Applying all these exclusions restricts the sample to 106,071 decisions, covering 412 judges across 53 court houses.

[5]Exceptions to the first-in-first-out rule occur when applicants are heard multiple times, file applications on additional issues, get delays, and 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.

Table 1
Asylum Judges: Summary Statistics

	Mean	Median	S.D.
Number of Judges	412		
Number of Courthouses	53		
Years Since Appointment	7.39	7.08	3.67
Daily Caseload of Judge	1.44	1.45	0.17
Average Grant	0.29	0.26	0.19
Moderate	33%		
Experienced	41%		
Lawyer	0.86		
Torture	0.02		
Defensive	0.56		
Family Size	1.27	1	1.06
Morning	51%		
Lunchtime	39%		

Table 1 summarizes our sample. The average years of experience among these judges is 7.4 years (we only have biographical data on 195 of the 412 judges, yet these 195 judges made 79,182 of the 106,071 decisions). The typical caseload of a judge is 1.44 asylum cases per day. Their average grant rate is 0.29. Moderate judges, defined as those whose average grant rate excluding decisions made on the day of the current decision is between 0.3 and 0.7, constitute 48,930 decisions. 41% of the judges have 9 or more years of experience. In the baseline sample of cases, 86% had a lawyer representing the applicant, 2% were withholding-convention against torture cases, and 56% were defensive cases meaning that the government initiated the case. The average family size is 1.27. 51% of decisions occurred in the morning between 8 AM and 12 PM, 39% occurred during lunch time between 12 PM and 2 PM, and 10% occurred in the afternoon from 2 PM to 8 PM. We mark the clock time according to the time that a hearing session opened.

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 litigant are treated as one decision as they tend to be all "grants" or all "denies". Multiple family members are treated as 1 observation for the same reason. Following Ramji-Nogales et al. (2007), we infer shared family status if cases share date, nationality, court, judge, decision, representation, and case type (affirmative or defensive).

Control variables in the regressions include, unless otherwise noted, a set of dummies for the number of yes 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 yes decisions over the past 5 decisions (excluding the current decision) across all judges in the court. This controls for recent trends in grants, case quality, or court mood. 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, torture indicator, defensive case indicator, family size, nationality fixed effects, and in some specifications, time of day (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 has been shown in Danziger et al., 2011).

4.3 Asylum Judges: Results

In Table 2, Column 1, we show that judges are 1.5 percentage points less likely to grant asylum to the current applicant if the previous decision an approve rather than a deny. In Column 2, we add controls for time-of-day fixed effects. This shows that our results are unlikely to be driven by other documented biases in judicial decision-making: variation in hunger or fatigue tied to time of day. Column 3 tests whether decisions are more likely to be reversed following streaks of previous decisions. The sample is restricted to decisions where the current and previous decisions both satisfy the requirement of occurring within one day or weekend after its previous decision. After a streak of two grants, judges are 2.1 percentage points less likely to grant asylum relative to decisions following a streak of two denials. Following a deny then grant decision, judges are 1.5 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, these magnitudes are economically significant: a 2 percentage point decline in the approval rate represents a 6.7% reduction in the probability of approval relative to the base rate of approval

of 29 percent.

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 applicant level. Observations are restricted to decisions that occurred within one day or weekend after the previous decision. Number of previous asylums granted is the full set of fixed effects for the number of grants out of the judge's previous 5 decisions (not including the current decision). It controls for the time-varying tendency of a judge to be positive in the recent time period. Number of previous asylums granted in the court is the full set of fixed effects for the number of grants within the 5 most recent cases in the same courthouse, excluding those of the judge corresponding to the current observation. Time of day controls consist of fixed effects for the start time of the hearing. Column 3 tests how judges react to streaks in past decisions. Lag Grant-Grant is a dummy equal to 1 if the judge approved the two most recent asylum applicants. Lag Deny-Grant is a dummy equal to 1 if the judge granted the most recent applicant and denied the applicant before that. Lag Grant-Deny is a dummy equal to 1 if the judge denied the most recent applicant and granted the applicant before that. The omitted category is Deny-Deny. The sample in Column 3 is restricted to decisions where the current and previous decision both satisfy the requirement of occurring within one day or weekend after its previous decision. Standard errors are clustered by judge. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Grant Asylum		
	(1)	(2)	(3)
Lag Grant	-0.0152***	-0.0149***	
	(0.00395)	(0.00395)	
Lag Grant - Grant			-0.0211**
			(0.00972)
Lag Deny - Grant			-0.0147**
			(0.00607)
Lag Grant - Deny			0.00183
			(0.00691)
Applicant Controls	Yes	Yes	Yes
Num Prev Asylums Granted by Judge	Yes	Yes	Yes
Num Prev Asylums Granted in Court	Yes	Yes	Yes
Time of Day	No	Yes	Yes
N	105983	105983	46455
R^2	0.199	0.200	0.197

Table 3, Column 1, shows that caseload exacerbates the negative autocorrelation in decisions. If there is only 1 case per day, there is no tendency to reverse the previous decision (because the sum of the coefficients on Lag Grant and Lag Grant x Caseload of Judge on that Day is approximately equal to zero). However, each additional case per day corresponds to an additional 2.7 percentage point reduction in the probability of approval if the previous case was approved. The average number of asylum cases that a judge handles per day is 1.44. Next, we show that the reduction

in the probability of approval following a previous grant is 4.4 percentage points greater when the previous decision is on the same day. Finally, Column 3 shows that the reduction in the probability of approval following a previous grant is 2.2 percentage points greater when the previous decision corresponds to an application with the same nationality as the current applicant. Altogether these results suggest 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 or was similar in terms of characteristics, i.e. nationality of the applicant.

Table 3
Asylum Judges: Heterogeneity by Court Characteristics

Column 1 tests how autocorrelation in decisions interacts with judicial case load. "Caseload" is the number of asylum cases decided by the judge on the day of the current decision. Column 2 tests whether the gambler's fallacy is stronger when the previous decision occurred on the same day as the current decision. Column 3 tests whether the gambler's fallacy is stronger when the previous decision concerned an applicant with the same nationality as the current applicant. 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.

	Grant Asylum		
	(1)	(2)	(3)
Lag Grant	0.0270***	0.000235	-0.0117***
	(0.00808)	(0.00414)	(0.00414)
Caseload of Judge on that Day	0.0157***		
	(0.00334)		
Lag Grant x Caseload of Judge on that Day	-0.0273***		
	(0.00493)		
Prev Decision is on Same Day		0.0266***	
		(0.00429)	
Lag Grant x Prev Decision is on Same Day		-0.0440***	
		(0.00708)	
Prev Decision is on Same Nationality			0.0346***
			(0.00552)
Lag Grant x Prev Decision is on Same Nationality			-0.0223*
			(0.0115)
Sample	All	All	All
Applicant Controls	Yes	Yes	Yes
Num Prev Asylums Granted by Judge	Yes	Yes	Yes
Num Prev Asylums Granted in Court	Yes	Yes	Yes
Time of Day	Yes	Yes	Yes
N	105983	105983	105983
R^2	0.201	0.201	0.201

Table 4, Column 1, shows that judges who are inexperienced (less than 10 years of experience) are

particularly likely to display negatively autocorrelated decisions. Autocorrelation in the decisions of experienced judges is statistically indistinguishable from zero. The effect becomes sharper when comparing within judges (Column 2 includes judge fixed effects) rather than cross-sectionally. Note that 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 Judge Experience remains informative. The sample is slightly smaller because we do not have access to biographies for all judges. Next we show that moderate judges (those with average grant rates of 30-70%, calculated excluding the day of the current decision) display stronger negative autocorrelation in decisions (Column 3).

Table 4
Asylum Judges: Heterogeneity by Judge Characteristics
This table tests how the negative autocorrelation decisions differs by judge experience and whether the judge is moderate. Experienced is defined as 10 or more years of experience. Moderate judges are those whose average grant rate outside of the current day is between 30% and 70%. 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.

	Grant Asylum		
	(1)	(2)	(3)
Lag Grant	-0.0158**	-0.0234***	0.00840
	(0.00664)	(0.00565)	(0.0108)
Experienced Judge	0.00200	0.00741	
	(0.0104)	(0.00846)	
Lag Grant x Experienced Judge	0.0197*	0.0388***	
	(0.0116)	(0.00930)	
Moderate Judge			0.111***
			(0.0113)
Lag Grant x Moderate Judge			-0.0379**
			(0.0186)
Sample	All	All	All
Applicant Controls	Yes	Yes	Yes
Num Prev Asylums Granted by Judge	Yes	Yes	Yes
Num Prev Asylums Granted in Court	Yes	Yes	Yes
Time of Day	Yes	Yes	Yes
Judge Fixed Effects	No	Yes	No
N	79123	79123	105983
R^2	0.196	0.239	0.209

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 experienced 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.

5 Loan Officers

5.1 Loan Officers: Data Description and Institutional Context

We use field experiment data collected by Cole et al. (2013).[6] 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 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

[6]For full details 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.

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). 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 empirical evidence suggests that loan offers still chose to reject many loans and may have experienced some other non-monetary or intrinsic 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 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 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 not 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.

5.2 Loan Officers: Empirical Specification Details

Observations are at the loan officer x loan order 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 to find gambler's fallacy 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 include a spline 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. 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 (as estimated using approval rates in other sessions). 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. 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 no 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 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 an equally-spaced 5-part spline for 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. 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 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 balance 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

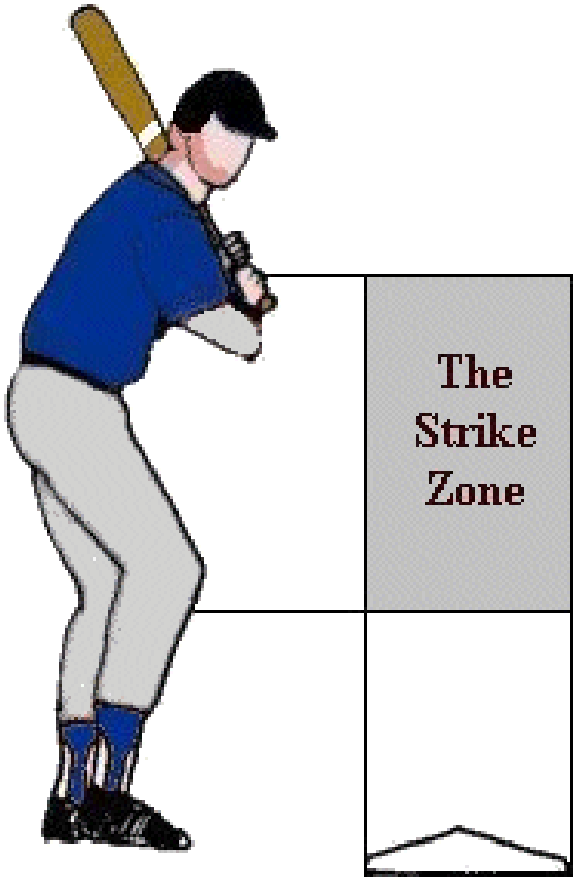
This table tests whether our results are robust to a balanced session design (each session consisted of 4 performing loans and 2 non-performing loans, randomly ordered). In Columns 1 and 2, we reproduce the results from Columns 1 and 3 of Table 6 showing that the negative autocorrelation in decisions 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). This sample restriction leaves us with approximately 1.5 million called pitches over 12,564 games by 127 different umpires.

Table 10
Baseball Umpires: Summary Statistics

Min. Year of Sample Coverage	2008
Max Year of Sample Coverage	2012
Number of Different Home Plate Umpires	127
Number of Games	12564
Average of Strike Dummy	0.3079
Fraction of Total Pitches Which Were Called Incorrectly	0.1282
Fraction of True Strikes Which Were Called Incorrectly	0.1733
Fraction of True Balls Which Were Called Incorrectly	0.1111
Fraction of Pitches Categorized as Ambiguous	0.1622
Fraction of Pitches Categorized as Obvious	0.3904
Fraction of Pitches Categorized as Neither	0.4474

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. 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.

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 (a dummy for each 3x3 inch square), a dummy 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. These controls 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.

Of course, umpires may be biased in other ways. For example, Moskowitz and Wertheim (2011) shows that umpires may act to avoid making calls that strongly determine game outcomes. To focus on the gambler's fallacy as distinct from these other biases, we control for dummies for every possible count (# 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 on the inning, score, outs, and number of players on base), score for the team at bat relative to the other team, and 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. 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 include the pitch location (a dummy for each 3x3 inch square), a dummy 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 (introduced as linear controls). We also control for dummies for every possible count combination (# balls and strikes called so far for the batter), the leverage index as defined previously, score of the team at bat relative to the other team, and whether the batter belongs to the home team.

In this experiment, we are particularly concerned that the "quality" 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 the distance of the pitch from the center of the strike zone. We estimate whether distance from the center of the strike zone depends on whether the lagged pitch was called a strike.

6.3 Baseball Umpires: Results

Table 11, Column 1, shows that umpires are 1 percentage point less likely to call a pitch a strike if the most recent previously called pitch was also 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, 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, then 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

	Full Sample		**Consecutive Pitches**	
Strike	(1)	(2)	(3)	(4)
Lag Strike	-0.00917^{***}	–	-0.0144^{***}	–
	(0.00059)		(0.00097)	
Lag Strike - Strike	–	-0.0131^{***}	–	-0.0207^{***}
		(0.00103)		(0.00268)
Lag Ball - Strike	–	-0.00995^{***}	–	-0.0187^{***}
		(0.000717)		(0.00156)
Lag Strike - Ball	–	-0.00271^{***}	–	-0.00666^{***}
		(0.000646)		(0.00155)
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 because

the batter did not swing). 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 significantly in this sample. Umpires are 2 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 all subsequent analysis, unless otherwise noted, we restrict the sample to consecutive called pitches.

Table 12
Baseball Umpires: Endogenous Pitcher Response

	No Location Controls		**With Location Controls**	
Distance from Center	(1)	(2)	(3)	(4)
Lag Strike	-0.022^{***}	–	0.000594	–
	(0.00226)		(0.00117)	
Lag Strike - Strike	–	-0.021^{**}	–	−0.00379
		(0.00772)		(0.0051)
Lag Ball - Strike	–	-0.0284^{***}	–	0.0016
		(0.00359)		(0.00193)
Lag Strike - Ball	–	-0.0102^{**}	–	0.000946
		(0.00326)		(0.0014)
N	891251	424706	891132	424645
R^2	0.0935	0.111	0.674	0.713

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 distance from the center of the strike zone as our dependent variable. To identify the effect of previous calls on pitch location, we exclude location controls in Columns 1 and 2. If pitchers are more likely to throw true balls after the previous pitch was called a strike, we should find significant positive coefficients. Instead we find significant negative coefficients. This implies that following a previous call of strike, the next pitch is likely to be closer to the center of the strike zone, which should make the next pitch more likely to be called a 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. In Columns 3 and 4, we estimate the same regressions 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

	Current Pitch Ambiguous		Current Pitch Obvious	
Strike	(1)	(2)	(3)	(4)
Lag Strike	-0.035^{***}	–	-0.00223^{***}	–
	(0.00377)		(0.000416)	
Lag Strike - Strike	–	-0.0484^{***}	–	-0.00514^{***}
		(0.0112)		(0.00101)
Lag Ball - Strike	–	-0.0332^{***}	–	-0.00439^{***}
		(0.00564)		(0.000772)
Lag Strike - Ball	–	-0.000827	–	-0.00279^{***}
		(0.0056)		(0.00084)
N	151501	73820	335318	153996
R^2	0.318	0.318	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 approximately ten times larger when the current pitch is ambiguous relative to when the current pitch is obvious. 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.

An important consideration that is specific to the baseball setting is that umpires may have a preference to be equally nice to two opposing teams and a desire to undo a previous marginal call. The desire to be equally nice to two opposing teams is unlikely to drive results in the asylum judge 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 to two opposing teams (or to be "fair" to the two teams) may, however, drive the negative autocorrelation of umpire calls within an baseball inning. After calling a marginal or difficult-to-call pitch a strike, the umpire may choose to balance out 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 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 probably 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.

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 $\beta_2 < 0$. Controlling for the 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 $Quality_{i,t-1}$.

Table 14 shows that sequential contrast effects are unlikely to drive our results in the case of asylum judges. In a first stage regression, we regress grant decisions on applicant controls. Using

the estimated coefficients, we create a predicted grant decision for each case, which functions as a continuous measure of the quality of the case. We then include the lagged case's predicted grant decision as a control variable. When we control for both the quality of the lagged case and actual lag decision, the current decision is negatively correlated with the previous decision, but significantly positively correlated the the continuous proxy for the previous case's quality. This is inconsistent with sequential contrast effects driving our results. The negative autocorrelation is stronger if the decision on the previous case was marginal rather than obvious.

Table 14
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 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 applicant was of higher quality, measured continuously. In other words, sequential contrast effects predicts that the coefficient on "Lag Grant Quality" should be negative. Applicant quality is measured as the predicted grant decision based on the asylum applicant's characteristics: legal representation, nationality, defensive or affirmative. Standard errors are clustered by judge. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Grant Asylum	
	(1)	(2)
Lag Grant	-0.0362***	-0.0362***
	(0.00476)	(0.00474)
Lag Case Quality	0.119***	0.120***
	(0.0137)	(0.0137)
Applicant Controls	Yes	Yes
Num Prev Asylums Granted by Judge	Yes	Yes
Num Prev Asylums Granted in City	Yes	Yes
Time of Day	No	Yes
N	105926	105926
R^2	0.0926	0.0938

Table 15 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 find that

the negative autocorrelation is stronger if the approval decision on the previous loan was marginal rather than obvious.

Table 15
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 a quota for the total number of positive decisions they can grant. Quotas considerations would also lead to negative autocorrelation in decisions. In all three of our empirical settings, agents do not face explicit quotas. For example, loan officers are only paid based upon accuracy and are explicitly told that they do not face quotas. However, one may be concerned about decision makers face 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 5, 10, or set of calls within a baseball inning that were called in a certain direction. We find that, conditional on these controls, extreme recency in the form of the previous single decision still negatively predicts the next decision.

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. The second related explanation is that agents are rational and understand random processes. However, agents prefer to alternate decisions over short time horizons, perhaps due to a desire to alternate being "mean" and "nice." This can be viewed as a preference for sequences that follow the law of small numbers rather than a belief in the law of small numbers. We cannot rule out this type preference for mixing entirely, but note that it is unlikely to drive behavior in the loan approval setting where loan officer decisions in the experiment do not affect real loan origination and subjects are paid purely for accuracy.

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 negatively 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. We estimate that the gambler's fallacy reverses up to 5% of all decisions. 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.

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.

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

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.