Decision-Making under the Gambler's Fallacy:

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

Daniel Chen	Tobias J. Moskowitz	Kelly Shue
ETH Zurich	University of Chicago and NBER	University of Chicago and NBER
Center for Law and Economics	Booth School of Business	Booth School of Business
chendan@ethz.ch	tobias.moskowitz@chicagobooth.edu	kelly.shue@chicagobooth.edu

March 11, 2015

Abstract

We find consistent evidence of negative autocorrelation in decision-making that is unrelated to the merits of the cases considered in three separate high-stakes settings: refugee asylum court decisions, loan application reviews, and baseball umpire calls. The evidence is most consistent with the law of small numbers and the gambler's fallacy – that people underestimate the likelihood of sequential streaks occurring by chance – leading to negatively autocorrelated decisions that result in errors. The negative autocorrelation is stronger among more moderate and less experienced decision-makers, following longer streaks of decisions in one direction, when the current and previous cases share similar characteristics or occur close in time, and when decision-makers face weaker incentives for accuracy. Alternative explanations for the negative autocorrelation such as sequential contrast effects, quotas, learning, or preferences to treat all parties fairly, are less consistent with the evidence.

*Most recent version at: https://sites.google.com/site/kellyshue/research/. We thank Dan Benjamin, John Campbell, Stefano Dellavigna, Andrea Frazzini, Emir Kamenica, Sendhil Mullainathan, Josh Schwartzstein, and Dick Thaler for helpful comments. We thank seminar participants at ANU, Cubist Systematic Strategies, Dartmouth, Econometric Society, Indiana University, International Society for New Institutional Economics, NBER Behavioral Economics, Rice, Rochester, SITE, University of Chicago, University of Oklahoma, University of Washington, and UNSW, for helpful comments. We also thank Alex Bennett, Luca Braghieri, Leland Bybee, Sarah Eichmeyer, Chattrin Laksanabunsong, and Kaushik Vasudevan for excellent research assistance and Sue Long for helpful discussions about the asylum court data.

1 Introduction

Does the sequencing of decisions matter for decision-making? Controlling for the quality and merits of each case, we find that the sequence of past decisions matters for the current decision: decision-makers exhibit negatively auto-correlated decision-making. Using three independent and high stakes settings: refugee asylum court decisions in the U.S., loan application reviews from a field experiment conducted in India by Cole, Kanz, and Klapper (2014), and Major League Baseball home plate umpire calls on pitches, we show consistent evidence of negatively autocorrelated decision-making, despite controlling for case quality, which leads to decision reversals and errors.

In each of the three high stakes settings, we show that the ordering of case quality is likely to be conditionally random. However, a significant percentage of decisions, more than five percent in some samples, are reversed or erroneous due to negative autocorrelation in the behavior of decision-makers. The three settings provide independent evidence of negatively autocorrelated decision-making across a wide variety of contexts. In addition, each setting offers unique advantages and limitations in terms of data analysis.

First, we test whether U.S. judges in refugee asylum cases are more likely to deny (grant) asylum after granting (denying) 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 U.S. 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 basis. 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 negative autocorrelation in underlying case quality. A limitation of the asylum court data, however, is that we cannot discern whether any individual decision is correct given the case merits. We estimate judges are up to 3.3 percentage points more likely to reject the current case if they approved the previous case. This translates to two percent of decisions being reversed purely due to the sequencing of past decisions, all else equal. This effect is also significantly stronger following a longer sequence of decisions in the same direction, when judges have "moderate" grant rates close to 50% (calculated excluding the current decision), and when the current and previous

cases share similar characteristics or occur close in time (which is suggestive of coarse thinking as in Mullainathan, Schwartzstein, and Shleifer, 2008). We also find that judge experience mitigates the negative autocorrelation in decision-making.

Second, we test whether loan officers are more likely to deny a loan application after approving the previous application using data from a loan officer field experiment conducted by Cole, Kanz, and Klapper (2014). The field experiment offers controlled conditions in which the order of loan files, and hence their quality, within each session is randomized by the experimenter. In addition, loan officers are randomly assigned to one of three incentive schemes, allowing us to 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. Another advantage of the field experiment setting is that payoffs 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 nine percent of decisions are reversed due to negative autocorrelation in decisions under the flat incentive scheme among moderate decision-makers, although the effect is significantly smaller under the stronger incentive schemes and among less moderate decision-makers. Across all incentive schemes, the negative autocorrelation is stronger following a streak of two decisions in the same 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 and vice versa. An advantage of the baseball umpire data is that it includes precise measures of the three-dimensional location of each pitch. Thus, while pitches may not be randomly ordered over time, we can control for each pitch's true "quality" or location and measure whether mistakes in calls conditional on a pitch's true location are 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 more than doubles when the current pitch is close to the edge of the strike zone (so it is a less obvious call) and is also significantly larger following two previous calls in the same direction. Put differently, MLB umpires call the same pitches in the exact same location differently depending solely on the sequence of previous calls. We also show

that any endogenous changes in pitch location over time are likely to be biases against our findings.

Altogether, we show that the existence of the negatively autocorrelated decision-making in three diverse settings is unrelated to the quality or merits of the cases considered and hence results in decision errors. Several potential explanations could be consistent with negatively autocorrelated decision-making, including the gambler's fallacy, sequential contrast effects, quotas, learning, and a desire to treat all parties fairly. We find that the evidence across all three settings is most consistent with the gambler's fallacy, and in several tests we are able to reject other theories.

The "law of small numbers" and the "gambler's fallacy" is the well documented tendency for 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) or underestimate the likelihood of streaks occurring by chance. For example, people often believe that a sequence of coin flips such as "HTHTH" is more likely to occur than "HHHHT" even though each sequence occurs with equal probability. Similarly, people may expect flips of a fair coin to generate high rates of alternation between heads and tails even though streaks of heads or tails often occur by chance. This misperception of random 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.

In our analysis of decision-making under uncertainty, a decision-maker who suffers from the gambler's fallacy may similarly believe that streaks of good or bad quality cases are unlikely to occur be chance. If so, the decision-maker will approach the next case with a *prior* belief that the case is likely to be positive if she deemed the previous case to be negative, and vice versa. Assuming that decisions made under uncertainty are at least partially influenced by the agent's priors, these priors then lead to negatively autocorrelated decisions. Similarly, a decision-maker who fully understands random processes may still engage in negatively autocorrelated decision-making in an attempt to appear fair if she is being evaluated by others, such as promotion committees or voters, who suffer from the gambler's fallacy.

Our analysis differs from the existing literature on the gambler's fallacy in two ways. First, most of the existing empirical literature examine behavior in gambling or laboratory settings (e.g. Benjamin, Moore, and Rabin, 2013; Ayton and Fischer, 2004; Croson and Sundali, 2005; Clotfelter and Cook, 1993; Terrell, 1994; Bar-Hillel and Wagenaar, 1991; Rapoport and Budescu, 1992; Jorgensen, Suetens, and Tyran, 2015; Asparouhova, Hertzel, and Lemmon, 2009). However, it has

not been shown whether the gambler's fallacy can bias high-stakes decision-making in real-world or field settings such as those involving judges, loan officers, and umpires.

Second, our analysis differs from the existing literature because we focus on decisions. We define a decision as the outcome of an inference problem using both a prediction and investigation of the current case's merits. In contrast, the existing literature on the gambler's fallacy typically focuses on predictions made by agents who do not also assess case merits. Our focus on decisions also highlights how 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. Our findings support this view across all three of our empirical settings.

Other potential alternative and perhaps complementary explanations appear less consistent with the data, though in some cases we cannot completely rule them out. The first is sequential contrast effects (SCE), in which decision-makers evaluate new information in contrast to what preceded it (Pepitone and DiNubile, 1976). Bhargava and Fisman (2014) 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, the decision-maker's criteria for quality while judging the current case is higher if the previous case was particularly high in quality. For example, after reading a great book, one's standard for judging the next book to be "good" or "bad" on a binary scale may be higher. Equivalently, the decision-maker's perception of the quality of the current case may be lower if the previous case was of very high quality. However, we show that SCE are unlikely to be a major driver of the negatively autocorrelated decisions in our three empirical settings. In both the asylum court and loan approval settings, we find that a decision-maker 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 binary decision. In the context of baseball pitches, there is an objective quality standard (the official regulated strike zone) that, in principle, should not move depending on the quality of the previous pitch.

Another potential alternative explanation is that decision-makers face quotas for the maximum number of affirmative decisions. Similarly, a learning model could also generate negatively autocorrelated decisions. In a learning model, decision-makers don't necessarily face quotas, but they believe that the correct fraction of affirmative decisions should be some level θ. The decision-makers

are unsure of where to set the quality bar to achieve that target rate and therefore learn over time, which could lead to negative autocorrelation in decisions.

However, in all three of our empirical settings, agents do not face explicit quotas or targets. For example, loan officers in the field experiment are only paid based upon accuracy and their decisions do not affect loan origination. Asylum judges are not subject to any explicit quotas or targets and neither are baseball umpires. Nevertheless, one may be concerned about self-imposed quotas or targets. We show that such self-imposed quotas are unlikely to explain our results by contrasting the fraction of recent decisions in one direction versus the sequence of such decisions. In a quotas model, the only thing that should matter is the fraction of affirmative decisions. We find, however, that agents negatively react to extreme recency holding the fraction of recent affirmative decisions constant. That is, if one of the last six or even two decisions was decided in the affirmative, it matters whether the affirmative decision occurred most recently or further back in time. This behavior is consistent with the gambler's fallacy. It is also largely inconsistent with self-imposed quotas, unless the decision-maker also has very limited memory and cannot remember beyond the most recent decision. Likewise, decision-makers in our settings are highly experienced and should have a standard of quality calibrated from many years of experience. They are probably not learning much from their most recent decision. Therefore, a learning model would not predict a strong negative reaction to the most recent decision, especially if we also control for their history of decisions.

A final potential interpretation issue that is specific to the baseball setting is that umpires may 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 in baseball may lead to negative autocorrelation of umpire calls if, after calling a marginal or difficult-to-call pitch a strike, the umpire chooses 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 (e.g., he, and everyone else, knew it was the right call to

make). Nevertheless, we find strong negative autocorrelation following these obvious and correct calls, suggesting that a desire to undo marginal calls or mistakes is not the sole driver of our results.

Lastly, we investigate two potential explanations closely related to the gambler's fallacy hypothesis. Instead of attempting to rule them out, we present them as possible variants of the same theme. 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. In other words, it is the outside monitors who have the gambler's fallacy and decision-makers merely cater to it. 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 loan officers know this) and the ordering of decisions 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 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).

Overall, we show that the gamblers fallacy and associated misperceptions of what constitutes a fair process can lead to decision reversals and errors. Consistent with experimental results in Gold and Hester (2008), in which the gambler's fallacy diminishes in coin flip predictions if the coin is allowed to "rest," we find that the negative autocorrelation in decisions declines if the current and previous case considered are separated by a greater time delay. Consistent with previous evidence showing that inexperience magnifies cognitive biases (Krosnick and Kinder, 1990; Berdejó and Chen, 2013), we find that education, experience, and strong incentives for accuracy can reduce biases in decisions. Our research also contributes to 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). Finally, our

results contribute to theoretical literature on judicial decision-making, e.g., Bordalo, Gennaioli, and Shleifer (2014), which models how judges can be biased by legally irrelevant information.

The rest of the paper is organized as follows. Section 2 outlines our empirical framework and discusses how it relates to theory. Section 3 presents the results for asylum judges. Section 4 presents results for the loan officer experiment. Section 5 presents the baseball umpire results. Section 6 discusses our findings in relation to various theories, with the evidence most consistent with the gambler's fallacy. Section 7 concludes.

2 Empirical Framework and Theory

We describe the general empirical framework we use to test for autocorrelation in sequential decision-making across the three empirical contexts and relate it to various theories of decision-making. In later sections when we describe each empirical setting in detail, we discuss how the empirical specifications are customized to fit the unique features of each setting.

2.1 Baseline Specification

Our baseline specification simply tests whether the current decision is correlated with the lagged decision, conditional on a set of control variables:

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

Y_{it} represents binary decisions by decision-maker i ordered by time t. β_1 measures the change in the probability of making an affirmative decision if the previous decision was affirmative rather than negative. If the ordering of cases is conditionally random, then β_1 should be zero if the quality of the case is the only determinant of decisions. $\beta_1 < 0$ is evidence in favor of negatively autocorrelated decision-making unrelated to quality, and $\beta_1 > 0$ is evidence of positive autocorrelation unrelated to quality. In some empirical settings, we can also determine whether any particular decision was a mistake. If we include a dummy for the correct decision as part of $Controls$, then any non-zero estimate of β_1 represents mistakes. In other settings when we cannot definitively determine a mistake, we use β_1 to estimate the fraction of decisions that are reversed due to autocorrelated

decision-making. For example, in the case of negative autocorrelation bias (what we find in the data), the reversal rate is: $-2\beta_1 a\,(1-a)$, where a represents the base rate of affirmative decisions in the data (see Appendix A for details).

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

We do not control for heterogeneity in the tendency to be positive 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 affirmative 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). Note, in constructing these and all other control variables, we never include the current observation in the calculation of averages, as that could lead to a mechanical negative estimated relationship between the current and previous decisions. 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 is not necessarily randomly ordered within each decision-maker. To attribute $\beta_1 < 0$ to decision biases, the underlying quality of the sequence of cases considered, conditional on the set of controls, should not itself be negatively autocorrelated. In the next sections, we discuss for each empirical setting why the sequences of cases appear to be conditionally random.[1]

Because many of our regressions include fixed effects (e.g., nationality of asylum applicant), we estimate all specifications using the linear probability model, allowing for clustered standard

[1] While we will present specific solutions to the possibility that case quality is not randomly ordered in later sections, we note that most types of non-random ordering are likely to correspond to persistent positive autocorrelation (e.g., slow-moving trends in refugee quality) which would bias against finding negative autocorrelation in decisions.

errors, as suggested by Angrist and Pischke (2008). However, we recognize that there is active ongoing debate in the econometrics literature concerning the relative merits of linear probability models versus logit or probit models when the dependent variable is binary. In unreported results, we reestimate all baseline tables using logit and probit models and find similar marginal effect estimates.

2.2 Moderate versus Extreme Decision-Makers

Throughout our analysis, we present additional results that distinguish between moderate and extreme decision-makers. We categorize an observation as corresponding to a moderate decision-maker if that decision-maker's average affirmative decision rate, calculated excluding the current decision (or current experimental session), is close to 0.5. We categorize a decision as corresponding to an extreme decision-maker if that decision-maker's average affirmative decision rate, calculated excluding the current decision (or current experimental session), is close to 0 or 1.

The designation of moderate versus extreme is not meant to imply that moderate decision-makers are more reasonable or accurate in their decisions. Rather, extreme decision-makers who tend to, for example, grant or deny asylum to all applicants cannot be very negatively autocorrelated in their decision process. In other words, extreme decision-makers mechanically cannot have β_1 very far below zero.[2] For full transparency, we present results for the full sample of decision-makers and for the sample of moderates.

It is important to note that restricting the regression sample to moderates does not mechanically generate $\beta_1 < 0$. This is because we never use the current observation in the calculation of whether a decision-maker is moderate. For example, suppose all decision-makers make decisions by flipping coins, so that the true autocorrelation in decisions is zero. Observations corresponding to these decision-makers would be categorized as moderates using their decision rates in other observations. However, we would still estimate that $\beta_1 = 0$, implying zero autocorrelation.

[2]Consider the regression $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$ for a sequence of decisions of a single decision-maker. Let $a \equiv P(Y = 1) = \beta_0/(1 - \beta_1)$ be the base rate of affirmatives in the data. Taking the conditional expectation of the regression yields $E[Y_t|Y_{t-1} = 1] = \beta_0 + \beta_1$ and $E[Y_t|Y_{t-1} = 0] = \beta_0$. Since both these quantities represent probabilities that are bound between 0 and 1, we have that $0 \leq \beta_0 + \beta_1 \leq 1$ and $0 \leq \beta_0 \leq 1$. When a is close to 1, $\beta_0 \to (1 - \beta_1)$. Since $\beta_0 \leq 1$, then β_1 cannot be negative. When a is close to 0, $\beta_0 \to 0$. Since $0 \leq \beta_0 + \beta_1$, then β_1 cannot be negative. Therefore, for extreme judges, for whom a is close to zero or one, β_1 cannot be too negative.

2.3 Relation to Theory

An autocorrelation coefficient, β_1, different from zero indicates that decision-makers are basing their decisions on something other than quality or satisfying an objective function that contains more than just accuracy. For instance, $\beta_1 > 0$ might imply some belief in the "hot hand," i.e., that seeing a recent streak of positive (or negative) cases implies something about the conditional quality of subsequent cases being higher (lower), even though the conditional quality has not changed. We find, however, that β_1 is significantly negative in all three of our settings, indicating that decision-makers are more likely to subsequently rule in the opposite direction after making a decision.

There are several theories that could be consistent with negatively autocorrelated decision-making. The first is the law of small numbers and the gambler's fallacy, which is the tendency for people to underestimate the probability of streaks occurring in random sequences. To motivate how the gambler's fallacy can lead to negatively autocorrelated decision-making, we present a simple extension of Rabin's (2002) model of the law of small numbers in Appendix B. The basic idea is that, if the ordering of cases is random and decisions are made only based upon case merits, a decision-maker's decision on the previous case should not predict her 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 decision-maker'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.

Another theory that can potentially deliver negatively autocorrelated decision-making is sequential contrast effects (SCE), which describe situations in which the decision-maker's criteria for quality while judging the current case is higher if the previous case was particularly high in quality (and vice versa). For example, after reading a great book, one's standard for judging the next book to be "good" or "bad" on a binary scale may be higher. Equivalently, the decision-maker's perception of the quality of the current case may be lower if the previous case was of very high quality. Like

the gambler's fallacy, SCE can lead to negative autocorrelation in binary decisions. Bhargava and Fisman (2014) find evidence of this in speed dating, for instance, where following a "very attractive" candidate, participants are more likely to reject the next candidate.

A third potential alternative explanation is that agents face quotas for the total number of affirmative decisions, which could lead to negative autocorrelation in decisions. In our three settings, decision-makers do not face any externally imposed quotas or recommendations. However, their decision-making could be influenced by self-imposed quotas. Related to a quotas model, a learning model could also generate negatively autocorrelated decisions. In a learning model, decision-makers don't necessarily face quotas, but they believe that the correct base rate of approve decisions should be some rate θ. The decision-makers are unsure of where to set the quality bar to achieve that target rate and therefore learn over time. So, for example, if a decision-maker just approved (too) many cases, she may decide that her quality bar is too low and raise it, to get to an overall approval rate of θ.

For all three of our empirical settings, we examine very high frequency autocorrelation in decision-making. In particular, we study whether agents react to the most recent decision. Given that the decision-makers in our settings have many years of experience, they are probably not learning much from their most recent decision and should have a standard of quality calibrated from many years of experience. As such, we do not believe a learning model best explains our findings. This mismatch with high-frequency autocorrelation also suggests that quotas are an unlikely explanation for our findings. Nevertheless, we explore these theories directly later on.

In what follows, we present our main empirical results, and for ease of exposition, discuss why the results are consistent with the law of small numbers and the gambler's fallacy. After presenting the empirical results, we then present additional tests exploring whether the results are also consistent with alternative explanations such as SCE, quotas, and learning. For the setting of baseball umpires, we also explore whether the negative autocorrelation could be caused by make-up calls, in which umpire attempt to be equally nice or fair to two opposing teams.

2.4 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:

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

All controls are as described in the baseline specification. 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. As we discuss later, tests for streaks can help differentiate among various theories. For example, a basic gambler's fallacy model predicts that $\beta_1 < \beta_2 < \beta_3 < 0$. The intuition is that agents mistakenly believe that streaks are unlikely to occur by chance, and longer streaks are particularly unlikely to occur. Following a (1,1) another 1 would constitute a streak of length three, which agents may believe is very unlikely to occur. Similarly, following a (0,1), agents may believe that another 1 is less likely to occur than a 0, because the former would create a streak of length two. Finally, following a (0,0) which is the omitted category, agents may believe a 1 is relatively more likely to occur because that would prevent a streak of 0's of length three.

The predictions under an SCE model are less obvious. If agents only contrast current case quality with the case that preceded it, then the decision in time $t-2$ should not matter, so we would expect $\beta_1 = \beta_2 < \beta_3 = 0$. However, if agents contrast the current case with the previous case and, to a lesser degree, the case before that, a SCE model could deliver similar predictions to those of the gamblers' fallacy model.

Conversely, a quotas model yields very different predictions. For quotas, β_1 should be the most negative, since two affirmative decisions in the past puts a more binding constraint on the quota limit than following only one affirmative decision. However, when the decision-maker decided in the affirmative for only one out of the two most recent cases, it should not matter whether the affirmative decision was most recent or not, hence $\beta_2 = \beta_3$. We test these various predictions across each of our three settings.

3 Asylum Judges

Our first empirical setting is U.S. asylum court decisions.

3.1 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 may reasonably fear imprisonment, torture, or death if forced to return to her home country. For a more detailed description of the asylum adjudication process in the U.S., we refer the interested reader to Ramji-Nogales et al. (2007).

We use administrative data on U.S. 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 his or her 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 in the previous case. To attribute negative autocorrelation in decisions to a cognitive bias, we first 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.[3] In other words, judges do not reshuffle the ordering of cases considered.

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) and therefore a bias against our findings of negative autocorrelation in decisions. 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). 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. Because many judges display extreme decision rates (close to zero or one), we also present subsample analysis excluding extreme judges or limited to moderate judges (grant rate close to 0.5). As discussed in Section 2.2, the extreme vs. moderate designation is not meant to imply that moderate judges are more accurate or reasonable. Rather, extreme decision-makers mechanically cannot have very negatively autocorrelated decisions. Further, because we exclude the current observation in the calculation of moderate status, our results within the moderate subsample will not spuriously generate findings of negative autocorrelation in the absence of true bias.

Judges are appointed by the Attorney General. 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 was $167,000. Based upon conversations with the

[3]Exceptions to the first-in-first-out rule occur when applicants file applications on additional issues 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). Since these violations of first-in-first-out are likely driven by applicant behaviors often several months prior to the recent set of decisions, they are likely uncorrelated with the judge's previous decision which often occurs in the same or previous day. In later tests, we also examine autocorrelation in proxies for case quality to assess whether deviations from the rule drive negative autocorrelation in decisions.

President of the National Association of Immigration Judges, no bonuses are granted. See Appendix C for more background information.

Our data comes from the Transactional Records Access Clearinghouse (TRAC). We exclude non-asylum related immigration decisions and focus on applications for asylum, asylum-withholding, or withholding-convention against torture. Applicants typically apply for all three types of asylum protection at the same time. 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 coded 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 TRAC data with our own hand-collected data on judicial biographies. 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 the 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 1,000 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

This table presents summary statistics of the asylum judges data that we use in our decision-making analysis.

	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 of asylum decisions. 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 approximately two 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.

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 0.2 and 0.8. The moderate indicator tags decisions for which the average grant rate for the judge for that nationality-defensive category, excluding the current observation, is between 0.3 and 0.7.[4]

3.2 Asylum Judges: Empirical Specification Details

Observations are at the judge × 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 on 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).[5] Observations in which there is a longer time gap between the current case and the lagged case are excluded from the sample (in unreported results, we find insignificant, close to zero, autocorrelation when the current and previous cases are separated by time lags of longer than one day).

Control variables in the regressions include, unless otherwise noted, a set of dummies for the number of affirmative decisions over the past five 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 five decisions across other judges (excluding the current judge) in the same court. This controls for recent trends in grants, case quality, or mood at the court level. 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 ×

[4]Results are not sensitive to using different, but similar ranges to define non-extreme and moderate judges.

[5]Results are similar if we exclude cases where the previous case occurred prior to the weekend.

defensive category, calculated excluding the current observation. We also control for the court's average grant rate for the relevant nationality × defensive category, calculated excluding the judge associated with the current observation. As noted previously, we don't include judge fixed effects because they mechanically induce 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 × defensive status 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).

3.3 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 approval rather than a denial. 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 to grant or deny asylum to almost all applicants from certain nationalities. As discussed previously, we do not claim in this paper that these more extreme judges necessarily make incorrect decisions. However, extreme judges necessarily exhibit less negative autocorrelation in their decisions. In Column 2 of Table 2, we restrict the sample to non-extreme judge observations. The extent of negative autocorrelation doubles to 1.1 percentage points. These results are evidence of negatively autocorrelated decision-making, consistent with decision-making under the gambler's fallacy. In unreported results, we find, as expected, that the autocorrelation in decisions among the omitted extreme judge observations sample is insignificant and close to zero.

Our sample consists of cases that follow another on the same day or in the previous working day. In Column 3 of Table 2, we further 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 results find near zero autocorrelation among sequential cases evaluated on different days. The fact that negative autocorrelation in decision-making is stronger when the two cases considered occur more closely in time is broadly consistent with the gambler's fallacy decision-making model, because

more recent cases may be more salient and lead to stronger expectations of reversals. These results are also consistent with experimental results in Gold and Hester (2008). They find that laboratory subjects who are asked to predict coin flips exhibited less gambler's fallacy after an interruption when the coin "rests." The higher saliency of more recent cases could also be consistent with stronger SCE, but is less likely to be consistent with a quotas constraint, unless judges self impose daily but not overnight quotas. Similarly, Column 4 restricts the sample to cases in which the current and previous case have the same defensive status. Individuals seeking asylum affirmatively, where the applicant initiates, are very different from those seeking asylum defensively, where the government initiates. The negative autocorrelation increases to 3.3 percentage points. (Conversely, among sequential cases with different defensive status, we find close to zero autocorrelation).

Finally, Column 5 of Table 2 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, whereas behavior following a grant then deny decision is insignificantly different from behavior following a streak of two denials. In the terms of our empirical framework introduced in Section 2.4, we are finding that $\beta_1 < \beta_2 < \beta_3$. These results are consistent with the gambler's fallacy affecting decisions and inconsistent with a basic quotas constraint model.

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 percent 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 1.6% of all decisions would have been reversed absent the negative autocorrelation in decision-making.

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 cases. *Lag deny-grant* is an indicator for whether the judge granted the most recent case and denied the case before that. *Lag grant-deny* is an indicator for whether the judge denied the most recent case and granted the case 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 across judges and cases. In this and subsequent tables, we restrict our analysis to the sample defined in Column 3 of Table 2 – observations for which the current and previous case were decided by non-extreme judges on the same day and with 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 nationality. This suggests that the negative autocorrelation in decisions may be tied to saliency and coarse thinking. Judges are more likely to engage in negatively autocorrelated decision-making when the previous

case considered is similar in terms of characteristics, in this case nationality. These results are consistent with stronger autocorrelation also found when the previous case occurred close in time with the current case or shared the same defensive status (as shown in the previous table).

In Column 2 of Table 3, we further show that moderate judges display stronger negative autocorrelation in decisions. Finally, Columns 3 and 4 show that judges who are inexperienced (less than the median experience in the sample – 8 years) display stronger negative autocorrelation. Experience is associated with significantly reduced negative autocorrelation both cross sectionally (Column 3) and within judges over time (Column 4).[6]

Because we measure decisions rather than predictions, reduced negative autocorrelation does not necessarily imply that experienced judges are more sophisticated in terms of understanding random processes. Both experienced and inexperienced judges could 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.

Finally, we present evidence supporting the validity of our analysis. To attribute negative autocorrelation in decisions to cognitive biases, we first 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 with that 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 $\times$ defensive status fixed effects. We estimate this regression using the entire sample of decisions, and create predicted grant status for each case using

[6]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* $\times$ *experienced judge* remains informative, which we focus on in Column 4.

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

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 rather than negatively autocorrelated, indicating that the negative autocorrelation we find in decision-making is not driven by case quality.

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 control variables 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 grants). Columns 3-5 use the same set of control variables as in Table 2, except that control variables for lawyer and family size are omitted because these are used as the dependent variables. 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

4 Loan Officers

Our second empirical setting examines loan officers making loan application decisions.

4.1 Loan Officers: Data Description and Institutional Context

We use field experiment data collected by Cole et al. (2014).[7] The original intent of the experiment was to explore how various incentive schemes affect the quality of loan officers' screening of loan applications. The framed field experiment was designed to closely match the underwriting process for unsecured small enterprise loans in India. Real 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

[7]For a detailed description of the data, we refer the interested reader to Cole et al. (2014). 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 serial correlation in decision-making were conducted. Therefore, differences between the subsample and full sample should not bias the analysis in favor of our findings.

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 at least one evaluation session. In each session, the loan officer screened six randomly ordered loan files and decided whether to approve or reject the loan file. Because the loan files corresponded 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 during 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). In our sample, the loan officers correctly classify loans approximately 65 percent of the time.

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, and $\overline{w}$ 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, loan officers might 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 approval rate and the ordering of decisions was never reported. Therefore, there was no incentive to negatively autocorrelate decisions for any reason.

In the "flat" incentive scheme, payoffs take the form $[20, 20, 0]$, so loan officers had monetary incentives to approve loans regardless of loan quality. However, as discussed above, loan officers may have had reputational concerns that led them to exert effort and reject low quality loan files even within the flat incentive scheme.[8] In the "stronger" incentive scheme, payouts took the

[8]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

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 an 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 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

This table presents summary statistics on the sample of loan officers obtained from Cole et al. (2014) that we use in our decision-making analysis.

	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 $\times$ loan file level. Overall, loan officers made correct decisions approximately 65% of the time. We consider an observation to correspond

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.

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, calculated excluding the current session, was closer to 0.5 than non-moderates.

4.2 Loan Officers: Empirical Specification Details

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 reliance on the previous decision to necessarily 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 mechanically induces negative correlation between Y_{it} and $Y_{i,t-1}$. Instead, to control for heterogeneity in mean approval rates at the loan officer $\times$ 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, to control for the fact that these type of loan officers are likely to have particularly high approval rates in the current session. 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).

4.3 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 sequencing of applications. 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 part of the loan

officer. After including this control variable, we find quantitatively similar results.

In Columns 3 and 4 of Table 6, 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). In the loan officers experimental setting, a potential additional reason 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). 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. 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. 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. Older and more experienced loan officers also display significantly reduced negative autocorrelation. These results are similar to our previous findings on asylum judges, and suggest that education, experience, and effort can reduce behavioral biases. Again, 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 cognitive biases such as the gambler's fallacy. These loan officers may still suffer equally 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 (misinformed and based-on-case-sequence) 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. The effects are larger and more significant when restricted to moderate judges (Column 2). Note that "Lag reject - approve" has a slightly less negative coefficient than "Lag approve - reject" even though a gambler's fallacy model where recency matters 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 (and vice versa). We believe there are several reasons why it is unlikely that loan officers would react to the balanced session design.

First, loan officers were never informed that sessions were balanced. Instead, they were told that the six loans within each session were randomly selected from a large population of loans. Second, if loan officers had "figured out" that sessions were balanced, we would expect that loan officers would be more likely to use this information when subject to stronger pay for performance. In other words, there should be greater negative autocorrelation within the incentive treatments with stronger pay-for-performance – this is the opposite of what we find. (Also, the better educated may be more likely to deduce that sessions are balanced, so they should display stronger negative autocorrelation, which is also opposite of what we find.)

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. This result is inconsistent with loan officers realizing that sessions were balanced. 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.

Finally, in unreported analysis, we find similar amounts of negative autocorrelation if we limit our sample to the first or last session attended by loan officers who participated in multiple sessions. This is inconsistent with a story in which, over time, loan officers learned that sessions were balanced and used that information in assessing loan quality. In addition, 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. If the strong negative autocorrelation in decisions under the flat incentive scheme is caused by loan officers realizing 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 results are unlikely to be generated by rational agents reacting to knowledge of 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

5 Baseball Umpires

Our final empirical setting uses data on called pitches by the home plate umpire in Major League Baseball (MLB).

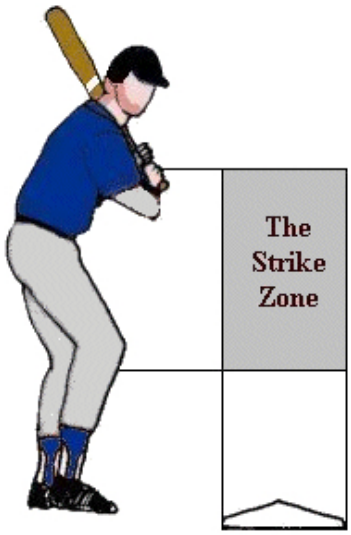
5.1 Baseball Umpires: Data Description and Institutional Context

In Major League Baseball, 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 described and 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. Hence, umpire judgment matters.

We test whether baseball umpires are more likely to call the current pitch a ball after calling the

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."



previous pitch a strike. Of course, pitch quality is not randomly ordered. For example, a pitcher will adjust his strategy depending on game conditions. An advantage of the baseball umpire data is that it includes precise measures of the trajectory and location of each pitch. Thus, while pitch quality 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 are negatively predicted by the previous call.

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 MLB 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 900 thousand observations.

Table 10
Baseball Umpires: Summary Statistics

This table presents summary statistics on the sample of umpire calls that we use in our decision-making analysis.

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 86.6% 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, controlling for pitch location, which should be the sole determinant of the call. To attribute negative autocorrelation in decisions to cognitive biases, 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. First, we control for the pitch location relative to an absolute point on home plate (indicators for each 3×3 inch square). We also control explicitly for whether the current pitch was within the strike zone based on its location, which should be the only characteristic that matters for the call according to MLB rules. Finally, we control for the speed, acceleration, curvature, and spin in the x, y, and z directions of the pitch, which may affect an umpire's perception. For a complete detailed

list of all control variables, please see Appendix D. Altogether, our 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. Nothing else, according to the rules, should matter for the call except the location of the pitch relative to the strike zone. Specifically, any coefficient on the lagged umpire decision will represent mistakes.

Of course, umpires may be biased in other ways. 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.[9] More relevant for our tests, Moskowitz and Wertheim (2011) show that umpires prefer to avoid making calls that result in terminal outcomes or may determine game outcomes. To differentiate our finding from these other types of biases which may affect the probability of the umpire calling strike versus ball at different points in the game, we control for indicator variables for every possible count combination (# balls and strikes called so far on the batter),[10] 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), 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.

Finally, note that our analysis uses the sample of called pitches, i.e. pitches in which the batter did not swing (so the umpire makes a call). Whether the batter chooses to swing is unlikely to be random and may depend on various game conditions. However, endogenous sample selection of this form should not bias our results toward finding spurious negative autocorrelation in umpire

[9] Along the same lines, umpires may potentially be misled by catcher framing, in which catchers strategically try to catch a pitch close to the chest, so that the pitch appears closer to the center of the strike zone than it actually was. In general, deceptive maneuvers such as catcher framing may alter the overall rate of called strikes within a game or inning, but should not affect our results which measure high-frequency negative autocorrelation. We test whether the current mistake in umpire decisions is negatively related to the previous call. Catcher framing should not affect negative autocorrelation in calls because catchers do not have incentives to frame *more* following a previous call of ball. In unreported tests, we also show that results remain very similar after the inclusion of catcher fixed effects, which account for the possibility that some catchers are better at framing than others.

[10] 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.

calls. We test, within the sample of called pitches, whether umpires tend to make mistakes in the opposite direction of the previous decision, after controlling for the true quality (location) of the current pitch. We also show in later analysis that, within the sample of called pitches in which the batter does not swing, insofar is pitch quality is not randomly ordered, it tends to be slightly positively autocorrelated, i.e., a bias against our findings of negative autocorrelation in umpire decisions.

5.2 Baseball Umpires: Empirical Specification Details

The sample includes all called pitches except for the first in each game or inning (because the previous pitch would be from a different game or different team). 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 Appendix D.

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. To address this concern, we control for each pitch's true location (plus the other controls described in Appendix D) and measure whether mistakes in calls conditional on a pitch's true location are negatively predicted by the previous call. We also re-estimate the regression by replacing the dependent variable of whether a pitch is *called* a strike with an indicator for whether the pitch was *actually* a true strike. 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 depend on whether the lagged pitch was called a strike. In other words, how does the actual quality of the pitch respond to the previous call? As shown in the next section, we find that changes in pitch quality are likely to bias us against finding negative autocorrelation in calls.

In unreported tests, we find that our results remain very similar with the inclusion of pitcher, batter, and umpire identity fixed effects. We don't include these controls in our main specifications because fixed differences across players should not bias estimates in favor of negative autocorrelation in calls, controlling for pitch location. For example, some pitchers may be more likely to throw strikes than others, but this should not lead to more negative autocorrelation in sequential umpire calls.

5.3 Baseball Umpires: Results

Table 11, Column 1 shows that umpires are 0.9 percentage points 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. These findings are consistent with the gambler's fallacy. These findings are less consistent with a quotas explanation (also umpires do not face explicit quotas). The results are also less consistent with a learning model about where to set a quality cutoff bar, because there is an objective quality bar (the official strike zone) that, according to the rules, should not move depending on the quality of the previous pitch.

All analysis in this and subsequent tables includes 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, and an indicator for whether the batter belongs to the home team. For a complete detailed list of control variables, please see Appendix D. 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 the consecutive 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 because that is the same baseline sample we use to estimate negative autocorrelation in umpire decisions. 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. 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 because the quality of pitches is slightly positively autocorrelated. 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 for whether our location controls account for variation in pitch location. All reported coefficients on lagged calls become small and insignificantly different from zero, indicating that our controls effectively remove any autocorrelation in the quality of pitches and account for pitcher's endogenous responses to previous calls.

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 the 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. Columns 1 and 2 restrict the analysis to observations in which the current pitch is ambiguous – pitches located close to the boundary of the strike zone, where it is difficult to make a correct strike or ball call. Columns 3 and 4, on the other hand, restrict the analysis to observations in which the current pitch is likely to be obvious – pitches located close to the center of the strike zone ("obvious" strikes) or far from the

edge of the strike zone ("obvious" balls). We find that the magnitude of negative autocorrelation 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 well below 0.001. This is consistent with the gambler's fallacy model that the decision-maker's prior beliefs about case quality will have less impact on the decision when the signal about current case quality is more informative.

It is also important to note that the stronger negative autocorrelation in the "current pitch ambiguous" sample is not an automatic consequence of our sample restriction. When the current pitch is difficult to call, we expect the raw umpire accuracy rate to decline. However, an unbiased umpire should not be more likely to make mistakes in the *opposite* direction of the previous call of strike or ball, which is what our stronger negatively autocorrelation coefficient shows.

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 on base. *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 negatively autocorrelated decision-making. A one standard deviation increase in umpire accuracy reduces negative autocorrelation by 25 percent. Finally, Column 3 tests whether the magnitude of the

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. The marginally stronger negative autocorrelation effects for high leverage situations and high attendance games may be consistent with umpires worrying about appearing biased in more heavily scrutinized environments, where fans, analysts, the media may suffer from the gambler's fallacy.

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-maker reviews 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 a 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 the negative autocorrelation is stronger following a previous correct call than following a previous incorrect call. This is inconsistent with a fairness motive, because umpires concerned with fairness should be more likely to reverse the previous call if it was incorrect. Column 2 shows that the negative autocorrelation remains equally strong or stronger when the previous call was obvious. 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 (e.g., he, and everyone else, knew it was the right call to make). 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 stronger following obvious calls (of which 99% are called correctly) and also 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.[11]

6 Addressing Alternative Explanations

We believe our results are best explained by decision-makers suffering from the gambler's fallacy. While we cannot completely rule out alternative explanations, we believe the alternatives have more difficulty reconciling all of the facts and offer some additional tests of these alternative stories.

[11] Another test we ran looked at the effect of the last called pitch for the previous team at bat on the first called pitch for the opposing team at bat. Fairness to the two teams would suggest that if an umpire called the pitch one way or made an error in one direction against one team, then he would make that same call on the opposing team to balance it out. This implies that there should be positive autocorrelation in calls or mistakes when the inning changes from one team at bat to the other team. We find no evidence consistent with this prediction.

6.1 Sequential Contrast Effects

Negative autocorrelation can arise if agents view the current case in contrast to the preceding case. For example, after reading a really great book, one's standard for judging the next book to be "good" or "bad" on a binary scale may be higher, leading to decision reversals. While SCE can be an important determinant of decision-making, we present a number of tests showing that SCE are unlikely to be a major driver of the negatively autocorrelated decisions we find in our three empirical settings.

To address this possibility, we 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 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 $Quality_{i,t-1}$.

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 the actual lagged decision, the current decision is negatively correlated with the previous decision, but not reliably related to the previous case's quality. We can reject that $\beta_2 < 0$. This is inconsistent with sequential contrast effects. We find $\beta_1 < 0$ and β_2 close to zero, which is consistent with a simple gambler's fallacy model such as that presented in Appendix B.[12]

[12]Under a simple model of the gambler's fallacy in decision-making, agents 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 decisions if they are more certain that the previous case was a true 1 (0) because it was very high (low) in quality. Such a model would also predict that $\beta_2 < 0$. Empirically, we find that β_2 is close to zero, contrary to the predictions of both the SCE model and this more nuanced model of 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 the standardized 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	0.00691*	0.00520
	(0.00385)	(0.00360)
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. During the experiment, the loan officers were asked to assess the quality of each loan application on a 100 point scale. These scores did not directly affect experiment payoffs, but Cole et al. (2014) show that 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 regress the current decision on the lagged binary decision and the lagged quality score. To adjust for the fact that some loan officers may be unobservably lenient, such that they tend to approve loans and assign high scores (leading to upward bias on β_2), we standardize the scores using the mean and standard deviation of each loan officer's reported scores in other sessions excluding the current session. We again find evidence contrary to the predictions of a sequential contrast model. We find $\beta_1 < 0$ and β_2 close to zero. This provides evidence from another independent setting that supports the gambler's fallacy model and is inconsistent with SCE.

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	-0.0223	-0.0736***
	(0.0148)	(0.0264)
Lag loan quality	0.00679	0.00692
	(0.00994)	(0.0201)
p-value lag loan quality rating < 0	0.247	0.365
Sample	All	Moderates
N	7495	2615
R^2	0.0252	0.0225

Finally, we cannot completely rule out SCE in baseball. As shown earlier in Table 15, umpires are slightly more likely to reverse the next call when the previous pitch was an obvious strike, i.e., high quality. Nonetheless, we note that SCE may simply be less likely to occur in this context because there is a well-defined quality metric: did the pitch fall inside or outside the regulated strike zone? In principle, for this setting, quality is defined and established by rule. However, there still may be room here for SCE to affect perceptions of quality, at least at the margin.

6.2 Quotas and Learning

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. Even without a self-imposed quota, decision-makers may believe that the correct fraction of affirmative decisions should be some level θ. Under a learning model, the decision-maker may be unsure of where to set the quality bar to achieve an affirmative target rate of θ, and learn over time. We show that self-imposed quotas or targets are unlikely to explain our results by controlling for the fraction of the previous two or five decisions that were made in a certain direction. We find that, controlling for the fraction of the previous five 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 decision, 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 decisions granted constant. This behavior is consistent with models of the gambler's fallacy. It is also largely inconsistent with self-imposed quotas, unless the decision-maker has limited memory and cannot remember beyond the most recent decision. Likewise, decision-makers in our settings are highly experienced and should have a standard of quality calibrated from many years of experience. They are probably not learning much from their most recent decision. Therefore, a learning model would not predict a strong negative reaction to the most recent decision, especially if we also control for their history of decisions using the fraction of recent decisions decided in the affirmative.

6.3 External Perceptions and Preferences for Alternation

Finally, we discuss two additional potential explanations for negatively-autocorrelated decisions that are closely related to our gambler's fallacy hypothesis. Instead of attempting to rule them out (which we cannot in any case), 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, even if they know they are wrong, 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 recent 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).

7 Conclusion

We document strong negative autocorrelation by decision-makers, unrelated to the quality of cases, in three high-stakes contexts: refugee asylum courts, loan application review, and baseball umpire calls. We find consistent evidence with many common links across the three independent settings. This negative autocorrelation is stronger among more moderate and less experienced decision-makers, following longer streaks of decisions in one direction, when the current and previous cases share similar characteristics or occur close in time, and when decision-makers face weaker incentives for accuracy. We show that the negative autocorrelation in decision-making is most consistent with the gambler's fallacy inducing decision-makers to erroneously alternate decisions because they mistakenly believe that streaks of affirmative or negative decisions are unlikely to occur by chance. We further show that the results are unlikely to be driven by potential alternative explanations such as sequential contrast effects, quotas, learning, or preferences to treat two teams fairly.

Beyond the three settings we study, the gambler's fallacy could effect decision-making more broadly. For example, financial auditors, HR interviewers, medical doctors, and policy makers all make sequences of decisions under substantial uncertainty. Our results suggest that misperceptions of what constitutes a fair process can perversely lead to unfair or incorrect decisions in many situations.

References

Angrist, Joshua D., and Jörn-Steffen Pischke, 2008, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press).

Asparouhova, Elena, Michael Hertzel, and Michael Lemmon, 2009, Inference from Streaks in Random Outcomes: Experimental Evidence on Beliefs in Regime Shifting and the Law of Small Numbers, *Management Science* 55, 1766–1782.

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.

Bar-Hillel, Maya, and Willem A Wagenaar, 1991, The Perception of Randomness, *Advances in Applied Mathematics* 12, 428–454.

Benjamin, Daniel, Don Moore, and Matthew Rabin, 2013, Misconceptions of Chance: Evidence from an Integrated Experiment, *Working Paper* .

Berdejó, Carlos, and Daniel L. Chen, 2013, Priming Ideology? Electoral Cycles without Electoral Incentives among U.S. Judges, *Working Paper* .

Bhargava, Saurabh, and Ray Fisman, 2014, Contrast Effects in Sequential Decisions: Evidence from Speed Dating, *Review of Economics and Statistics* 96, 444–457.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, 2014, Salience Theory of Judicial Decisions, *Journal of Legal Studies* Forthcoming.

Clotfelter, Charles T., and Philip J. Cook, 1993, The "Gambler's Fallacy" in Lottery Play, *Management Science* 39, 1521–1525.

Cole, Shawn, Martin Kanz, and Leora Klapper, 2014, Incentivizing Calculated Risk-Taking: Evidence from an Experiment with Commercial Bank Loan Officers, *Journal of Finance* Forthcoming.

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.

Executive Office for Immigration Review, 2014, Office of the Chief Immigration Judge, "http://www.justice.gov/eoir/ocijinfo.htm".

Gold, E., and G. Hester, 2008, The Gambler's Fallacy and a Coin's Memory, in Joachim I. Krueger, ed., *Rationality and Social Responsibility: Essays in Honor of Robyn Mason Dawes*, 21–46 (Psychology Press, New York).

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

Jorgensen, Claus Bjorn, Sigrid Suetens, and Jean-Robert Tyran, 2015, Predicting Lotto Numbers, *Journal of the European Economic Association* Forthcoming.

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.

Pepitone, Albert, and Mark DiNubile, 1976, Contrast Effects in Judgments of Crime Severity and the Punishment of Criminal Violators, *Journal of Personality and Social Psychology* 33, 448–459.

Political Asylum Immigration Representation Project, 2014, *Appearing at a Master Calendar Hearing in Immigration Court*, 98 North Washington Street, Ste. 106, Boston MA 02114.

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.

Rapoport, Amnon, and David V. Budescu, 1992, Generation of Random Series in Two-Person Strictly Competitive Games, *Journal of Experimental Psychology: General* 121, 352–363.

Terrell, Dek, 1994, A Test of the Gambler's Fallacy—Evidence from Pari-Mutuel Games, *Journal of Risk and Uncertainty* 8, 309–317.

TRAC Immigration, 2008, Improving the Immigration Courts: Effort to Hire More Judges Falls Short, "http://trac.syr.edu/immigration/reports/189/".

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 A: Calculation of Reversal Rate

Consider the simple regression $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$. Taking expectations, $P(Y = 1) = \beta_0 / (1 - \beta_1)$. Let $a \equiv \beta_0 / (1 - \beta_1)$ be the base rate of affirmatives in the data. Suppose that, absent the bias toward negative autocorrelation in decisions, the average approval rate would still equal a. If the previous decision was a negative, then the negative autocorrelation 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 negative autocorrelation is $(\beta_0 - a) \cdot P(Y_{i,t-1} = 0) + (a - (\beta_0 + \beta_1)) \cdot P(Y_{i,t-1} = 1)$. To simplify, substitute $\beta_0 = a(1 - \beta_1)$, so that the previous equation simplifies to $-2\beta_1 a(1 - a)$, which is positive since $\beta_1 < 0$.

Appendix B: A Model of Decision-Making Under the Gambler's Fallacy

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 the gambler's fallacy and belief in the law of small numbers. In the Rabin model, agents who suffer from the gambler's fallacy 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 biased agent 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 behavioral agent'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.

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).[13]

[13] In this simple model of the gambler's fallacy in decision-making, agents form priors based upon previous decisions. In a more nuanced model of the gambler's fallacy, along the lines of the model in Rabin and Vayanos (2010), agents may react more negatively to previous decisions if they are more certain that the previous decision was correct. Such a model would yield similar predictions to those of a SCE model in which agents are more likely to reverse previous decisions if the previous case was very low or high in quality, measured continuously. As shown in Section 6.1, we do not find strong empirical evidence of agents reacting negatively to continuous quality measures of the previous case,

The agent also observes an i.i.d. 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 an agent who suffers from the gambler's fallacy. 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.

Following Rabin (2002), we assume an agent who suffers from the gambler's fallacy 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 belief in the law of small numbers is indexed by $N \in \mathbb{N}$ and we assume $N \geq 6$. As $N \to \infty$, the biased agent behaves likes the rational thinker.

Model Predictions

The simple model generates the following testable predictions for decision-makers who suffer from the gambler's fallacy:

1. Decisions will be negatively autocorrelated as long as the signal of case quality is not perfectly informative. This occurs because decisions depend on prior beliefs which are negatively related to the previous decision.

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. This follows immediately from Rabin (2002).

3. The negative autocorrelation will be stronger following a streak of two or more decisions in the same direction. This follows from an extension of Rabin (2002) where the decision-maker

after conditioning on the previous binary decision. In other words, the empirical results are more consistent with the simple gambler's fallacy model presented in this section.

believes that he is making the first, second, or third draw from the urn, each with probability one-third.

4. The negative autocorrelation in decisions is stronger when the signal about the quality of the current case is less informative. This follows directly from the threshold decision rule defined above.

Appendix C: Additional Background on Asylum Judges

Immigration Courts Overview

The immigration judges are part of the Executive Office for Immigration Review (EOIR), an agency of the Department of Justice (Political Asylum Immigration Representation Project, 2014). At present, there are over 260 immigration judges in 59 immigration courts. In removal proceedings, immigration judges determine whether an individual from a foreign country (an alien) should be allowed to enter or remain in the United States or should be removed. Immigration judges are responsible for conducting formal court proceedings and act independently in deciding the matters before them. They also have jurisdiction to consider various forms of relief from removal. In a typical removal proceeding, the immigration judge may decide whether an alien is removable (formerly called deportable) or inadmissible under the law, then may consider whether that alien may avoid removal by accepting voluntary departure or by qualifying for asylum, cancellation of removal, adjustment of status, protection under the United Nations Convention Against Torture, or other forms of relief (Executive Office for Immigration Review, 2014).

Immigration Judges

The immigration judges are attorneys appointed by the Attorney General as administrative judges. 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. See INA sec. 101(b)(4) (8 U.S.C. 1101(b)(4)); 8 CFR 1003.10(b), (d). Decisions of the immigration judges are subject to review by the Board pursuant to 8 CFR 1003.1(a)(1) and (d)(1); in turn, the Board's decisions can be reviewed by the Attorney General, as provided in 8 CFR 1003.1(g) and (h). Decisions of the Board and the Attorney General are subject to judicial review (Executive Office for Immigration Review, 2014).

In our own data collection of immigration judge biographies, many previously worked as immigration lawyers or at the Immigration and Naturalization Service (INS) for some time before they were appointed. The average tenure of active immigration judges, as of 2007, was approximately eleven to twelve years. Since 2003 the annual attrition rate has averaged approximately 5%, with the majority of departures due to retirement (TRAC Immigration, 2008).

Proceedings before Immigration Courts

There are two ways an applicant arrives to the Immigration Court. First, the asylum seeker can affirmatively seek asylum by filing an application. In the event that the Asylum Office did not

grant the asylum application[14] and referred it to Immigration Court, the asylum seeker can now pursue his or her asylum claim as a defense to removal in Immigration Court. Second, if the asylum seeker never filed for asylum with the Asylum Office but rather the government started removal proceedings against him or her for some other reason, he or she can now pursue an asylum case in Immigration Court (Political Asylum Immigration Representation Project, 2014). This latter group is classified as defensive applicants and includes defendants picked up in immigration raids.

Families

We treat multiple family members as a single case because family members almost always receive the same asylum decision (based upon Ramji-Nogales et al. (2007) and verified from conversations with several asylum judges). 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. (2007) 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.

A potential concern with inferring that two applicants 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 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 and strong negative autocorrelation when the current and previous case do not correspond to the same nationality. This type of negative autocorrelation is extremely unlikely to be generated by errors in the inference of families because family members will almost always have the same nationality.

Appendix D: MLB Control Variables

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

[14] For application at the Asylum Office, see chapters 14-26 of: http://immigrationequality.org/get-legal-help/our-legal-resources/immigration-equality-asylum-manual/preface-and-acknowledgements/

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