The Quarterly Journal of Economics Advance Access published March 24, 2016

DECISION-MAKING UNDER THE GAMBLER'S FALLACY: EVIDENCE FROM ASYLUM JUDGES, LOAN OFFICERS, AND BASEBALL UMPIRES *

Daniel Chen Tobias J. Moskowitz Kelly Shue

January 12, 2016

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 field settings: refugee asylum court decisions, loan application reviews, and major league baseball umpire pitch calls. The evidence is most consistent with the law of small numbers and the gambler's fallacy – people underestimating 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. Other explanations for negatively autocorrelated decisions such as quotas, learning, or preferences to treat all parties fairly, are less consistent with the evidence, though we cannot completely rule out sequential contrast effects as an alternative explanation. (JEL codes: D03, G02, D8)

*Corresponding author: Kelly Shue, University of Chicago and NBER, 5807 S Woodlawn Ave, Chicago, IL, 60601, (734) 834-0046, kelly.shue@chicagobooth.edu. We thank Dan Benjamin, John Campbell, Kent Daniel, Stefano Dellavigna, Andrea Frazzini, Radha Gopalan, Emir Kamenica, Adrien Matray, Sendhil Mullainathan, Josh Schwartzstein, Dick Thaler, Jeff Zwiebel, three anonymous referees, and Andrei Shleifer (the editor) for helpful comments and suggestions. We thank seminar participants at AEA, ANU, Conference on Behavioral and Experimental Economics, Conference on Empirical Legal Studies, Cornell, Cubist, Dartmouth, Econometric Society, Gerzensee ESSFM, Indiana University, ISNIE, McGill, NBER Behavioral Economics, Northeastern, Norwegian School of Economics, Red Rock Finance Conference, Rice, Rochester, SITE, Texas Finance Festival, University of Chicago, University of Oklahoma, University of Washington, UNSW, Yale Summer School in Behavioral Finance, and Zurich 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 a case, we find that the sequence of past decisions matters for the current decision – decision-makers exhibit negatively autocorrelated decision-making. Using three independent and high stakes field settings: refugee asylum court decisions in the U.S., loan application reviews from a field experiment by Cole, Kanz, and Klapper (2015), 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 induced by the behavior of decision-makers. The three settings provide independent evidence of negatively autocorrelated decision-making across a wide variety of contexts for decision-makers in their primary occupations, and across a very large sample size of decisions in some cases. Each field setting offers unique advantages and limitations in terms of data analysis that taken together portray a compelling picture of negatively autocorrelated decision-making arising from belief biases.

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 on 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 into two percent of decisions being reversed purely due to the sequencing of past decisions, all else equal. This effect is also stronger follow-

ing 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 in India by Cole, Kanz, and Klapper (2015). 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 incentives reduce 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 decision-making under the flat incentive scheme among moderate decision-makers. The effect is significantly weaker 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. Education, age, experience, and a longer period of time spent reviewing the current loan application reduce 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, holding pitch location fixed. 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 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. We explore several potential explanations that could be consistent with negatively autocorrelated decision-making, including belief biases such as the gambler's fallacy and sequential contrast effects, and other explanations such as 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/or sequential contrast effects, and in several tests we are able to reject the 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 can lead to errors in predictions.

In our analysis of decision-making under uncertainty, a decision-maker who himself suffers from the gambler's fallacy may similarly believe that streaks of good or bad quality cases are unlikely to occur by chance. Consequently, the decision-maker may 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 several ways. First, most of the existing empirical literature examines 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; Suetens, Galbo-Jorgensen, and Tyran, 2015; Asparouhova, Hertzel, and Lemmon, 2009) and does not test 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 professional baseball umpires.[1]

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 or bets made by agents who do not also assess case merits. Our focus on decisions 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.

Finally, we study the behavior of experienced decision-makers making decisions in their primary occupations. In some settings, we have variation in incentives to be accurate and show that stronger incentives can reduce the influence of decision biases on decisions. In addition, and in contrast to the laboratory setting as well as other empirical settings studied in the literature, our decision-makers see a large sample of cases – many hundreds for an asylum judge and tens of thousands or more for an umpire – affording us very large samples of decisions.

Other potential alternative and perhaps complementary explanations appear less consistent with the data, though in some cases we cannot completely rule them out. One potential alternative explanation is that decision-makers face quotas for the maximum number of affirmative decisions, which could induce negative autocorrelation in decisions since a previous affirmative decision implies fewer affirmative decisions can be made in the future. 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

[1]Simonsohn and Gino (2013) (SG) also examine decisions in a real-world setting by looking at the scoring of MBA admissions interviews. They focus on narrow bracketing (dividing continuous flows of judgments into daily subsets), although they discuss the gambler's fallacy as a potential mechanism behind their findings. While SG examine scores on a 1-5 scale, we study binary sequences of decisions, which may be a closer fit to simple binary models of the gambler's fallacy. In addition, and importantly, we emphasize difference in reactions to the ordering of recent decisions while SG test a general narrow bracketing model in which agents react to the average score assigned previously within the same day, regardless of ordering. As we highlight, the sequencing and ordering of cases is a key distinguishing feature of the gambler's fallacy.

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 with 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 N 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 sequencing of decisions mattering and is largely inconsistent with self-imposed quotas, unless the decision-maker also has very limited memory and cannot remember beyond the most recent decision.

Another related potential explanation is a learning model, where decision-makers do not necessarily face quotas, but they believe that the correct fraction of affirmative decisions should be some level. 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, baseball umpires should not have a target rate and instead have a quality bar (the official strike zone) that is set for them. Further, decision-makers in all of our settings are highly experienced and should therefore have a standard of quality calibrated from many years of experience. As a consequence, they are probably not learning much from their most recent decision or sequence of decisions. In addition, a learning model would not predict a strong negative reaction to the most recent decision either, especially when we also control for their own recent history of decisions, which should be a better proxy for learning.

Another potential interpretation specific to the baseball setting is that umpires may have a preference to be equally "fair" to both 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 "make it up" to the team at bat 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.

Finally, we investigate several potential explanations closely related to the gambler's fallacy. Since these are empirically indistinguishable, we present them as possible variants of the same theme, though we argue they may be less plausible in some of our settings. The first is sequential contrast effects (SCE), in which the decision-maker's perception of the quality of the current case is negatively biased by the quality of the previous case (Pepitone and DiNubile, 1976; Simonsohn and Loewenstein, 2006; Simonsohn, 2006). For example, Bhargava and Fisman (2014) find that speed dating subjects are more likely to reject the next candidate if the previous candidate was very attractive and Hartzmark and Shue (2015) find that investors perceive today's earnings news as less impressive if unrelated firms released good earnings news in the previous day. Theoretically, the gambler's fallacy and SCE can predict the same patterns in decision outcomes. The distinction is mainly with regard to *when* the subject makes a quality assessment. Under the gambler's fallacy, a subject who sees a high quality case will predict that the next case is likely to lower in quality in a probabilistic sense even before seeing the next case, whereas SCE predicts the subject will make a relative comparison after seeing both both cases. While the laboratory or prediction markets may be able to separate these two biases, they will be observationally equivalent when looking at only decision outcomes, since we cannot observe what is inside a decision-maker's head. Complicating matters further, it may also be the case that the gambler's fallacy affects the decision-maker's perception of quality, leading to a contrast effect. For example, a subject may believe that the next case is likely to be lower in quality after seeing a high quality case, and this makes him perceive the next case as indeed being less attractive.

We present suggestive evidence that our results are more consistent with a simple gambler's fallacy model than the SCE model. SCE may be less likely to occur in the context of baseball because there is a well-defined quality metric (the regulated strike zone), although SCE may still bias perceptions of quality on the margin. In both the asylum court and loan approval settings, we find that decisions are unrelated to continuous quality measures of the previous case after we

condition on the previous binary decision. This is consistent with a simple gambler's fallacy model in which agents expect binary reversals and less supportive of a SCE model in which agents should react negatively to the continuous quality of the previous case. However, our tests cannot fully reject SCE because we may measure the true quality of the previous case with error.

Another possibility 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. While concerns about external perceptions could be an important driver of decisions, 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.

Lastly, a related explanation is that agents may prefer to alternate being "mean" and "nice" over short time horizons. This preference could, again, originate from the gambler's fallacy. A decision-maker who desires to be fair may over-infer that she is becoming too negative from a short sequence of "mean" decisions. However, a preference to alternate mean and nice is unlikely to drive behavior in the loan approval setting where loan officers 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 belief biases possibly stemming from misperceptions of what constitutes a fair process can lead to decision reversals and errors. While we cannot completely distinguish between variants of the gambler's fallacy and SCE, our evidence is unique to the literature on decision-making biases in its breadth in terms of studying large samples of important decisions made as part of the decision-maker's primary occupation. We also find heterogeneity in the field data that may have useful policy implications. For example, we find that negative autocorrelation in decisions declines if the current and previous case considered are separated by a greater time delay, 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 further find that education, experience, and strong incentives for accuracy can reduce biases in decisions. Finally, our research also contributes to the sizable psychology literature using vignette studies of small samples of judges that suggest unconscious heuristics (e.g., anchoring, status quo bias, availability) play a role in judicial decision-

making (e.g., Guthrie et al., 2000). In addition, our results contribute to the theoretical literature on 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, including the gambler's fallacy. Section 7 concludes.

2 Empirical Framework and Theory

We describe our empirical framework for testing autocorrelation in sequential decision-making across the three empirical contexts and relate it to various theories of decision-making.

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. 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. $\beta_1 < 0$ is evidence in favor of negatively autocorrelated decision-making unrelated to quality, and $\beta_1 > 0$ is evidence of positive autocorrelation unrelated to quality. 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.[2] $\beta_1 < 0$ could be consistent with several theories, including the gambler's fallacy, which we show through 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 more negative if she deemed the previous case to be positive, and vice versa, leading to negatively autocorrelated decisions. Negative autocorrelation in decisions could also be consistent with sequential contrast effects (SCE), quotas, and learning.

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 is evidence of 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, since the previous and current decision are both positively correlated with the decision-maker's unobserved tendency to be positive.

We control for decision-maker heterogeneity in several ways. One simple method is to control for heterogeneity using decision-maker fixed effects. However, decision-maker fixed effects within a finite panel can lead to negative correlation between any two decisions by the same decision-maker, which biases toward $\beta_1 < 0$. To remove this bias, we alternatively control for a moving average of the previous n decisions made by each decision-maker, not including the current decision. A benefit

[2]Following Gilovich et al. (1985)'s seminal work, a number of papers have found evidence of hot hand beliefs in sports settings, although some results have been challenged in recent work, e.g., Miller and Sanjurjo (2014) and Green and Zwiebel (2015).

of this specification is that it also tests whether the decision-maker reacts more to the most recent decision, controlling for the average affirmative rate among a set of recent decisions. The drawback of using a moving average is that it may imprecisely measure the tendency of each decision-maker to be positive due to small samples and hence be an inadequate control for heterogeneity. Hence, we also control for the decision-maker's average decision in all other settings other than the current decision.[3] In our baseline results, we report estimates that control for individual heterogeneity using recent moving averages and leave-out-means because these methods do not bias toward $\beta_1 < 0$. In the Online Appendix, we show that the results are very similar with the inclusion of decision-maker fixed effects, although point estimates tend to be more negative, as expected. Finally, we cluster standard errors by decision-maker or decision-maker×session as noted.

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. We discuss for each empirical setting why the sequences of cases appear to be conditionally random.[4]

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 errors, as suggested by Angrist and Pischke (2008). However, we recognize there is debate in the econometrics literature concerning the relative merits of various binary dependent variable models. In the Online Appendix, we reestimate all baseline tables using logit and probit models and estimate similar marginal effects.

[3]Except for the regressions with decision-maker fixed effects, we never include the current observation in the calculation of averages for control variables, since that could lead to a spurious negative estimated relationship between the current and previous decisions in finite panels.

[4]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 positive autocorrelation (e.g., slow-moving trends in refugee quality) which would bias against finding negative autocorrelation in decisions.

2.2 Streaks

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

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

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. 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.

The predictions under an SCE model are less obvious and depend on the specific assumptions of the model. For instance, 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, implying $\beta_1 < \beta_2 < \beta_3 < 0$.

A quotas model, on the other hand, 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 $\beta_2 = \beta_3$. The learning model also does not predict $\beta_2 < \beta_3$ unless it is a particular form of learning where more weight is given to the most recent decision. 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. For these reasons, a judge may treat these cases differently or, at the very least, categorize them separately. Therefore, we also test whether the negative autocorrelation in decision-making is stronger when consecutive cases have the same defensive status (both affirmative or both defensive).[5]

The court proceeding at the immigration court level is adversarial and typically lasts several hours. Asylum seekers may be represented by an attorney at their own expense. A DHS attorney cross-examines the asylum applicant and argues before the judge that asylum is not warranted. Those that are denied asylum are ordered deported. Decisions to grant or deny asylum made by judges at the immigration court level are typically binding, although applicants may further appeal to the Board of Immigration Appeals.

Our baseline tests explore whether judges are less likely to grant asylum after granting asylum in the previous case. To attribute negative autocorrelation in decisions to a cognitive bias, we first

[5]See http://www.uscis.gov/humanitarian/refugees-asylum/asylum/obtaining-asylum-united-states for more details regarding the asylum application process and defensive vs. affirmative applications.

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.[6] 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 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 and test autocorrelation in observable proxies of case quality in the Online Appendix.

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 limiting to moderate judges (grant rate close to 0.5). We exclude the current observation in the calculation of moderate status, so 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

[6]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., closures may occur if the applicant doesn't show up, if the applicant chooses to withdraw, or for miscellaneous rare reasons encoded in the "other" category). 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. To test this, we also examine autocorrelation in proxies for case quality in the Online Appendix to assess whether deviations from the rule drive negative autocorrelation in decisions. We find nothing in this regard.

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

Our data comes from a FOIA request filed through the Transactional Records Access Clearinghouse (TRAC). We exclude non-asylum related immigration decisions and focus on applications for asylum, withholding of removal, or protection under the convention against torture (CAT). Applicants typically apply for all three types of asylum protection at the same time. As in Ramji-Nogales et al. (2007), when an individual has multiple decisions on the same day on these three applications, we use the decision on the asylum application because a grant of asylum allows the applicant all the benefits of a grant of withholding of removal or protection under the withholding-convention against torture while the reverse does not hold. In the Online Appendix we redefine a grant of asylum as affirmative if any of the three applications are granted and find qualitatively similar results.[7] 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 immediate prior decision by the judge is on the same day or previous day or over the weekend if it is a Monday. Finally, we restrict the sample to judges who review a minimum of 100 cases for a given court and courts with a minimum of 1,000 cases in the data. These exclusions restrict the sample to 150,357 decisions, across 357 judges and 45 court houses.

Table I summarizes our sample of asylum decisions. Judges have long tenures, with a median of

[7]Following Ramji-Nogales et al. (2007), we use the decision on the asylum application for our baseline analysis. If the judge denies asylum but grants withholding of removal or protection under CAT, the asylum applicant receives much more limited benefits than she would if she were granted asylum. In such cases, applicants face employment limitations and are only granted withholding of removal to the particular country where they may be persecuted but may be moved to a safe third country (and such protections are person-specific rather than applying to spouses or children). Therefore, it is not obvious whether a denial of asylum accompanied by a grant of withholding or protection under CAT is a positive or negative decision. Further, while the evidentiary standard for qualifying for withholding of removal or protection under CAT is much higher than those for the asylum application, the judge also exercises less subjective discretion in the determination of the former two applications which are classified as mandatory if the applicants meet the high evidentiary standard. This is relevant for cases in which the applicant has committed crimes or assisted in the persecution of others (which disqualify her for asylum) but remains eligible for withholding of removal or protection under CAT.

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 case load of a judge is approximately two asylum cases per day. The average grant rate is 0.29. 94% of cases have a lawyer representing the applicant, and 44% are defensive cases initiated by the government. The average family size is 1.21. 47% of hearings occur in the morning between 8 AM and 12 PM, 38% occur 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.[8]

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 regression 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), and for which we know the ordering of cases considered within the same day.[9]

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 grant 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 leave-out-mean grant rate for the relevant nationality × 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

[8] Results, reported in the Online Appendix, are qualitatively similar using these two sets of cutoffs.

[9] We also have data on decisions in which we do not know the ordering of the current case with respect to the previous case because two or more cases are considered within a single session with a single time stamp. These observations are excluded from the regression sample, but are used to create control variables relating to judge average grant rates.

associated with the current observation. In our baseline results, we do not include judge fixed effects because they mechanically induce a small degree of negative correlation between Y_{it} and $Y_{i,t-1}$. In the Online Appendix we report results using judge fixed effects and obtain similar results with slightly more negative coefficient estimates, as expected. Finally, we control for the characteristics of the current case: presence-of-lawyer indicator, family size, nationality $\times$ 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 II, 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, all else equal. In the remaining columns, we focus on cumulative subsamples in which the magnitude of the negative autocorrelation increases substantially. First, the asylum data cover a large number of judges who tend to grant or deny asylum to almost all applicants from certain nationalities. More extreme judges necessarily exhibit less negative autocorrelation in their decisions. In Column 2 of Table II, we restrict the sample to non-extreme judge observations (where non-extreme is calculated excluding the current decision). The extent of negative autocorrelation doubles to 1.1 percentage points.

In Column 3 of Table II, we further restrict the sample to cases that follow another case on the same day (rather than the previous day). We find stronger negative autocorrelation within same-day cases. The stronger negative autocorrelation when two consecutive cases occur more closely in time is broadly consistent with saliency and 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), which finds that laboratory subjects who are asked to predict coin flips exhibit 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 or multi-day quotas.

Column 4 of Table II restricts the sample further to cases in which the current and previous

case have the same defensive status. Individuals seeking asylum affirmatively, where the applicant initiates, can be very different from those seeking asylum defensively, where the government initiates. In affirmative cases, applicants typically enter the country legally and are applying to extend their stay. In defensive cases, applicants often have entered illegally and have been detained at the border or caught subsequently. Judges may view these scenarios to be qualitatively different. The negative autocorrelation increases to 3.3 percentage points.

Hence, from an unconditional 0.5 percentage points, the negative autocorrelation increases six-fold to 3.3 percentage points if we examine moderate judges on same-day cases with the same defensive status. Using the estimate in Column 4 of Table II within the sample of non-extreme, same-day, same defensive cases, the coefficient implies that 1.6% of asylum decisions would have been reversed absent the negative autocorrelation in decision-making. Table A.I in the Online Appendix reports the extent of the negative autocorrelation among each omitted sample and presents formal statistical tests for whether the estimates in each cumulative subsample significantly differ from one-another. We find that the negative autocorrelation among extreme-judge and different-defensive-status subsamples are close to zero and significantly differ from the non-omitted samples. However, the negative autocorrelation is economically substantial even across consecutive days (with insignificant differences), although the effect size doubles when the judge considers cases within the same day.

Finally, Column 5 of Table II 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 asylum 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 2.2, we find that $\beta_1 < \beta_2 < \beta_3 < 0$. A formal statistical test of the difference in the β's appears at the bottom of the table, where we reject that the betas are all equal and that $\beta_2 = \beta_3$. (The only insignificant difference is between β_1 and β_2, though β_1 has, as predicted, a more negative point estimate.) These results are consistent with the gambler's fallacy affecting decisions and inconsistent with a basic quotas model. Moreover, the magnitudes are economically significant. 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.

We report robustness tests of our findings in the Online Appendix. Table A.II reports results using logit and probit models. The economic magnitudes are similar. Table A.III reports results using judge fixed effects. As expected, the coefficients are slightly more negative due to a mechanical negative autocorrelation between any two decisions by the same judge induced by the fixed effects. However, the bias appears to be small and the coefficient estimates are of similar magnitude to our baseline results that control for a moving average of each judge's past five decisions as well as her leave-out-mean grant rate. In addition, the precision of the estimates do not change much between the two specifications, suggesting that controlling for heterogeneity using the moving average of a judge's decisions and her leave-out-mean instead of judge fixed effects yields similar identification despite the former containing more measurement error. Table A.IV presents results for an alternative definition of the granting of asylum, where instead of using the asylum grant decision, we code a decision as a grant if the judge granted any of the asylum, withholding of removal, or protection under the U.S. Convention Against Torture applications. The results are very consistent with slightly smaller point estimates.

Finally, a potential concern with the sample split among moderate and extreme decision-makers is that we may mechanically measure stronger negative autocorrelation among moderates. We emphasize that, because we do not use the current observation in the calculation of whether a decision-maker is moderate, restricting the sample to moderates does not mechanically generate $\beta_1 < 0$ if the true autocorrelation is zero (for example, a judge who decides based upon random coin flips would be classified as a moderate, but would display zero autocorrelation). However, another potential issue that could mechanically generate greater measured negative autocorrelation for moderate judges is our use of a binary statistical model. The autocorrelation of a binary variable is biased away from one and the size of the bias increases as the base rate of decisions gets closer to zero or one. This may lead us to estimate a lower degree for negative autocorrelation for "extreme" decision-makers. One method to address this issue is to use the tetrachoric correlation, which models binary variables as functions of continuous (bivariate normal) latent variables. The bivariate probit model extends the tetrachoric correlation to allow for additional control variables. Using the bivariate probit model, Table A.V shows that there is strong negative autocorrelation

in decisions for moderate decision-makers that is statistically significant and of similar economic magnitude as those from our baseline regressions. Conversely, for extreme decision-makers, there is no evidence of any autocorrelation. These results match our estimates from the linear probability, logit, and probit regressions and indicate that the potential mechanical correlation coming from binary models is not driving our results.

Table III explores additional heterogeneity across judges and cases. In this and subsequent tables, we restrict our analysis to the sample defined in Column 4 of Table II – 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 of Table III 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 Table II).

Columns 2 and 3 of Table III show that moderate judges and judges with less experience display stronger negative autocorrelation in decisions. Judges who have less than the median experience in the sample (8 years) display stronger negative autocorrelation. The fourth column repeats the regression including judge fixed effects. We find that experience is also associated with significantly less negatively autocorrelated decisions for a given judge over time.[10]

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

[10]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 except in tables in the Online Appendix because judge fixed effects bias the coefficient on *Lag grant* downward. However, the coefficient on *Lag grant* × *experienced judge* remains informative, which we focus on in Column 4.

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 and not case quality, we 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 support this assumption in Table A.VI in the Online Appendix. We find that case quality does not appear to be negatively autocorrelated in terms of observable proxies for quality. However, our identifying assumption requires that autocorrelation in unobserved aspects of case quality is also not negative.

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. (2015).[11] 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 processed, loan applications. Each loan file contained all the information available to the bank at the time the loan was first evaluated.

[11] For a detailed description of the data, we refer the interested reader to Cole et al. (2015). Our data sample consists of a subset of the data described in their paper. This 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.

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 application. 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 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, loan officers correctly classify loans approximately 65 percent of the time. The percentage of performing loans they approve is 78%, while the percentage of non-performing loans they approve is 62%, which shows they exhibit some ability to sort loans. Overall, the tetrachoric correlation between the binary variables, loan approval and loan performance (1 = performing, 0 = non-performing), is 0.29 and significantly different from random chance.

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, 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.[12] In the "stronger" incentive scheme, payouts take the form $[20, 0, 10]$, so loan officers faced a monetary incentive to reject non-performing loans. In the "strongest" incentive scheme, payouts take 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 IV 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. We consider an observation to correspond to a moderate loan officer if the average approval rate of loans by the loan officer in other sessions (not including the current session) within the same incentive scheme is between 0.3 and 0.7.

4.2 Loan Officers: Empirical Specification Details

Y_{it} is an indicator for whether the loan is approved. Loans are ordered within a session. Our sample includes observations for 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.

[12] The incentives in the "flat" scheme may at first seem surprisingly weak, but the authors of the original experiment used this incentive condition to mimic the relatively weak incentives faced by real loan officers in India. As shown in the next table, the overall approval rate within the flat incentive scheme is only 10 percentage points higher than the approval rates under the two other incentive schemes and loan officers were still more likely to approve performing than nonperforming loans. This suggests that loan offers still chose to reject many loans and may have experienced some other intrinsic or reputational motivation to accurately screen loans.

We control for heterogeneity in mean approval rates at the loan officer × incentive scheme level using 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 types 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). Because the loan officer field experiment data is limited in size and each session consists of only six loan decisions, we do not control for a moving average of each loan officer's average decision rate over the past five decisions within the session (as we do in the asylum judge setting). In the Online Appendix, we also present results controlling for loan officer fixed effects.

4.3 Loan Officers: Results

Table V, 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, as shown by the other interaction coefficients in Column 1. 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, indicating that the negatively autocorrelated decision-making results in decision errors.

In Columns 3 and 4 of Table V, we repeat the analysis for 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. Comparing the coefficient estimates 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.

Tables A.VII - A.X in the Online Appendix report robustness tests for the loan officer sample. Table A.VII further tests whether the loan officers in the experiment are exerting effort in making accurate decisions and how that effort varies with incentives. We assess whether the loan approval decision is correlated with the ex ante quality of the loan file, as proxied by the fraction of other loan officers who approved the loan file, and the average quality score/rating given by other loan officers for the loan file. We further explore how the correlation between decisions and ex ante loan quality interact with the three incentive schemes. The results show that loan officers are more likely to approve loans that other loan officers approve or rate highly and that the consensus in decision-making increases with stronger incentives.

Table A.VIII shows that the results are similar using other binary regression models, such as logit and probit. Table A.IX reports results from a specification that includes loan officer fixed effects. The coefficients are directionally similar but more negative than those in Table V. This is expected because the inclusion of fixed effects, particularly in short panels such as in the loan officers experimental setting, biases the coefficients downward. Table A.X reports results from a bivariate probit model that adjusts for the bias that, when using a binary dependent variable, the moderate subsample may mechanically exhibit more negative autocorrelation. The results using the bivariate probit model confirm that the negative autocorrelation in decisions is stronger for moderates even after adjusting for this potential bias, and the negative autocorrelation decreases with incentives.

In the remaining analysis, we pool the sample across all three incentive treatments unless otherwise noted. Table VI 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.[13] 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. 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 VII examines decisions following streaks of decisions. 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 loan officers (Column 2). We easily reject that $\beta_1 = \beta_2 = \beta_3$ and $\beta_1 = \beta_3$ for the past sequence of decisions. However, "Lag reject - approve" has a 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 and small in the sample of moderates.

Lastly, we 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.[14] 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.

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

[14]Note that the fraction of loans performing is not exactly 67%, implying that the original experiment did not implement an exactly balanced session design for every session. Cole, Kanz, and Klapper (2015) initially balanced each session to have exactly four performing loans, according to early performance data given to them by the bank that originally processed the loans. However, the bank then sent the researchers a revised categorization of the loan files. The researchers used the revised data to categorize the data but did not reassign loans to sessions. This led to 85% of the sessions in our data having exactly 4 performing loans, 11% of the sessions having 3 performing loans, 3% of the sessions having 5 performing loans, and 1% of sessions having 2 performing loans.

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 again the opposite of what we find.

In Columns 1 and 2 of Table A.XI in the Online Appendix, we reproduce the baseline results showing that the negative autocorrelation in decisions is strongest in the flat incentive scheme treatment. In Columns 3 through 6, we show that the true performance of the current loan is negatively related to both the lagged decision and the true quality of the lagged loan file, and the negative autocorrelation in true loan quality is approximately similar in magnitude across all three incentive treatments. The results in Columns 1 and 2 are 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.

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 home plate 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 passes over home plate is within the strike zone as described and shown in Figure I. 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 I, 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 previous pitch a strike. Of course, pitch quality (e.g., location) 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 crosses 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 and implemented for part of the 2007 season. Our data covers approximately 3.5 million pitches over the 2008 to 2012 MLB seasons, when the system produced an entire season of pitch data. We restrict our analysis to called pitches, i.e., pitches in which the batter does not swing (so the umpire must make a call), excluding the first called pitch in each 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, where 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.

Baseball umpires in our sample do not receive immediate feedback regarding whether each call was correct (data on whether each pitch was within the strike zone according to the PITCHf/x system is available after the game). Nevertheless, the umpire likely receives some cues. At the very least, umpires can observe the extent to which others disagreed with the call. First, the umpire knows roughly where the pitch landed and how "close" the call was. A call made on a pitch near the edge of the strike zone is more ambiguous, for instance. More to the point, the umpire also receives cues from the batter's reaction, the pitcher's reaction, and the crowd's reaction to the call. These parties voice their disagreement if they believe the umpire made a mistake. Making an unambiguous erroneous call will likely draw a stronger reaction from at least one of these parties.

Table VIII 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 fall within ±1.5 inches of the edge of the strike zone. 60% of ambiguous pitches are called correctly. Obvious pitches fall within 3 inches around the center of the strike zone or 6 inches or more outside the edge of the strike zone. 99% of obvious pitches are called correctly.

5.2 Baseball Umpires: Empirical Specification

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

To attribute negative autocorrelation in decisions to cognitive biases, we assume that the underlying quality of the pitches (e.g., 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 using 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. 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 significant coefficient on the lagged umpire decision is evidence of 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

differ in race and ethnicity. However, while biases against teams or specific types of players could affect the base rate of called pitches within innings or against certain pitchers, they should not generate high-frequency negative autocorrelation in calls, which is the bias we focus on in this paper.[15] In addition, in the Online Appendix, we include umpire, batter, and pitcher fixed effects, which should account for these sorts of biases, and find similar effects. More relevant for our tests, Moskowitz and Wertheim (2011) show that umpires prefer to avoid making calls that result in terminal outcomes or that 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 (number of balls and strikes called so far on the batter),[16] 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.

In our previous analysis of asylum judges and loan officers, we controlled for heterogeneity in each decision-maker's approval rate using the decision-maker's leave-out-mean approval rate, moving average of the past five decisions, and/or decision-maker fixed effects. We also conducted subsample analysis limited to moderate decision-makers. These control variables for decision-maker heterogeneity are less relevant in the setting of baseball umpires, because professional umpires tend to have very homogeneous mean rates of strike calls.[17] Therefore, we present our baseline regression results without inclusion of controls for individual heterogeneity (the lack of controls should be a bias against findings of negative autocorrelation), and show in the Online Appendix that the results are very similar if we control for umpire fixed effects or a moving average of the past five decisions.

[15] 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.

[16] In Table A.XVI in the Online Appendix, we find qualitatively similar coefficients on $Y_{i,t-1}$ if we do not control for count.

[17] Among umpires who have made more than 500 calls, the standard deviation in the mean rate of calling strikes is 0.01, potentially because extreme umpires would be much less accurate and umpire accuracy can be judged relative to the PITCHf/x system.

Finally, since we use the sample of called pitches, these are pitches in which the batter chose not to swing. Whether the batter chooses to swing is unlikely to be random and may depend on various game conditions, which is partly why we add all of the controls above. However, endogenous sample selection of this form should also not bias our results toward finding spurious negative autocorrelation in umpire 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 that, insofar as pitch quality is not randomly ordered, it tends to be slightly positively autocorrelated within this sample, which is a bias against our findings of negative autocorrelation.

5.3 Baseball Umpires: Results

Table IX, 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. We easily reject that $\beta_1 = \beta_2 = \beta_3$ in favor of $\beta_1 < \beta_2 < \beta_3 < 0$. These findings are consistent with the gambler's fallacy and less consistent with a quotas explanation (in addition, 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).

In Columns 3 and 4 of Table IX, 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) without any interruption. 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. 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.

Tables A.XII - A.XVI in the Online Appendix report robustness tests of these results. Table A.XII shows similar results with batter, pitcher, and umpire fixed effects. Table A.XIII reports similar results using the moving average of the umpire's past five calls as a control variable. Table A.XIV shows similar effects and economic magnitudes using logit and probit models, and Table A.XV shows similar results using a bivariate probit model. Table A.XVI shows similar results if we exclude control variables for the count (number of balls and strikes called so far on the batter).

Since in this setting, we are particularly concerned that the "quality", i.e., location, of the pitch will also react to the umpire's previous call, we 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. If our location and other controls are mismeasured or inadequate, however, then autocorrelation in the quality of pitches could still be an issue. To assess how concerning this issue might be, 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 is *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 then 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?

Table A.XVII in the Online Appendix shows that the negative autocorrelation in umpire calls is unlikely to be caused by changes in the actual location of the pitch. We continue to restrict the sample to consecutive called pitches and repeat the analysis using 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). Columns 1 and 2, which use an indicator for whether the current

pitch was within the strike zone as the dependent variable, show that pitchers are more likely to throw another strike after the previous pitch was called a strike, resulting in positive, rather than negative, coefficients on the previous call. Hence, autocorrelation in the quality of pitches biases us against our finding of negatively autocorrelated decision-making. 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; again, we find the opposite. 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 include the same set of detailed pitch location controls (dummies for each 3 x 3 inch square) as in our baseline specifications, and find that all coefficients on lagged calls become small and insignificant, suggesting that our controls effectively remove any autocorrelation in the quality of pitches and account for pitcher's endogenous responses to previous calls.

Table X 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, 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. We can confidently reject equality of the estimates for ambiguous and obvious pitches in Columns 1 and 3 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 for ambiguous pitches is not merely a consequence of these pitches being more difficult to call. We expect umpire accuracy to decline for these pitches, but an unbiased umpire should not be more likely to make mistakes in the *opposite* direction from the previous call. That is, overall accuracy may be lower but there is

no expectation that calls should alternate and be negatively autocorrelated.

In Table XI, 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 a 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) are also 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, and the media may suffer from the gambler's fallacy.

6 Addressing Alternative Explanations

Across all three of our settings, we find consistent evidence of negatively autocorrelated decision-making. We believe our results are best explained by decision-makers suffering from the gambler's fallacy. Other explanations for negatively autocorrelated decisions such as quotas, learning, or preferences to treat all parties fairly, are less consistent with the evidence, though we cannot completely rule out sequential contrast effects as an alternative explanation.

6.1 Sequential Contrast Effects

Perhaps the most difficult alternative story to distinguish – and one which we will not be able to fully reject empirically – is sequential contrast effects (SCE). Under SCE, negative autocorrelation

in decisions can arise if agents view the current case in contrast to the preceding case. SCE imply that lagged case quality affects the perception of the quality of the current case. Under Rabin's (2002) original model, where agents react to past binary outcomes, we could, in principle, distinguish between agents responding to past decisions (the gambler's fallacy) versus lagged quality (SCE), where the former is a binary outcome and the latter continuous. Specifically, we could estimate:

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

If SCE drives our findings, then we expect to find that $\beta_2 < 0$, holding constant the previous discrete decision $Y_{i,t-1}$. The idea is, holding constant the previous discrete decision, SCE predicts that decision-makers should be more likely to reject the current case if the previous case was of high quality, as measured continuously using $Quality_{i,t-1}$. However, in a more general model of the gambler's fallacy, such as that proposed in Rabin and Vayanos (2010), agents may react more negatively to the previous decision 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 $\beta_2 < 0$, and hence the two theories make identical predictions.

Tables A.XVIII and A.XIX in the Online Appendix estimate the above equation for the asylum judges and loan officers samples, respectively, and find that, using a continuous predicted quality measure for asylum cases and loan officer's quality scores for loans, the current decision is negatively correlated with the previous decision, but not reliably related to the previous case's quality. This is consistent with a simple gambler's fallacy model as in Rabin (2002) and less consistent with SCE or a more general model of the gambler's fallacy in which agents react negatively to the continuous quality of the previous case. However, the test cannot fully reject SCE because we may measure the true quality of the previous case with error. If unobserved quality is better captured by the binary decision rather than the observed continuous quality measure, then both coefficients represent quality and are consistent with both SCE and the gambler's fallacy.

Likewise, we cannot completely rule out SCE in baseball. In principle, SCE may simply be less likely to occur in the context of baseball because there is a well-defined quality metric: the regulated strike zone. Therefore, quality is established by rule. However, there still may be room for SCE to affect perceptions of quality, at least at the margin. Further, as shown later in Table XII, umpires

are slightly more likely to reverse the next call when the previous pitch was an obvious strike, i.e., high quality, which is consistent with SCE.

Theoretically, the main distinction between the gambler's fallacy and SCE lies in *when* the subject makes a quality assessment. Under the gambler's fallacy, a decision-maker who just reviewed a high quality case would predict the next case is less likely to be high quality (because two high quality cases in a row are unlikely to occur) even before seeing the next case, whereas, under SCE, the decision-maker will make a relative comparison after seeing both cases. While the laboratory setting may be able to separate these two biases, they are observationally equivalent when looking at only decision outcomes, since we cannot observe what is inside a decision-maker's head. Complicating matters further, contrast effects bias could potentially arise from the gambler's fallacy. After reviewing a high quality case and deciding in the affirmative, a decision-maker may believe that the next case is less likely to be of high quality, and this makes her perceive the next case as indeed having lower quality, resulting in a contrast effect.

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 self-impose 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 N decisions that were made in the affirmative, where N equals 2 or 5, and testing whether the previous single decision still matters. We find that, holding constant the fraction of the previous two or five decisions decided in the affirmative, the previous single decision negatively predicts the next decision (see Tables II and IX and A.XIII). The only exception is the loan officers setting in which we do not find, controlling for the fraction of the past two decisions made in the affirmative, that loan officers react more negatively to the most recent decision. However, the results are less precisely estimated because the

field experiment data offers a shorter panel and smaller sample size.[18] In general, this behavior is consistent with models of the gambler's fallacy, and largely inconsistent with self-imposed quotas, unless the decision-maker has very limited memory and cannot remember beyond the most recent decision. Likewise, decision-makers in our three 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 recent decisions using the fraction of recent decisions decided in the affirmative. In addition, baseball umpires should make decisions according to an objective quality standard (the officially defined-strike zone) rather than according to a target affirmative decision rate.

6.3 External Perceptions and Preferences for Alternation and Fairness

Finally, we discuss two additional possible explanations for negatively-autocorrelated decisions that are closely related to variants of our gambler's fallacy 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 autocorrelated decisions, even if they know they are wrong, in order to avoid the appearance of being too lenient or too harsh. 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 are 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

[18] While we cannot show that loan officers react negatively to the most recent decision controlling for the fraction of recent decisions made in the affirmative, other results appear inconsistent with a quotas or learning explanation. The loan officers are paid for accuracy and they should be more likely to self-impose quotas or learn how to implement a target decision rate when they face stronger monetary incentives.

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).

An important and related 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. 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. However, in baseball, the umpire makes sequential calls on the same team at bat. Fairness motives may lead umpires to undo a previous marginal or mistaken call, which could result in negative autocorrelation. 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 do not seek to completely rule out these types of situations, we show that "make-up" calls and preferences for fairness appear unlikely to drive our estimates for baseball umpires.[19]

In Table XII, Column 1 shows that the negative autocorrelation is stronger following a previous correct call than following a previous incorrect call, which 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. These results suggest that fairness concerns and a desire

[19]We also tested 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 two teams would suggest that, if an umpire called a 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 positive autocorrelation in calls when the inning changes. We find no evidence consistent with this prediction.

to be equally nice to two opposing teams are unlikely to explain our results.

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 reviews, and professional 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 cannot rule out that sequential contrast effects also help to explain to these findings, but we show that the results are unlikely to be driven by other alternative explanations such as quotas, learning, or preferences to treat parties fairly.

Beyond the three settings we study, negatively autocorrelated decision-making could have broader implications. For example, financial auditors, human resource 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.

Appendix A: Calculation of Reversal and Mistake Rates

In this section, we discuss how to interpret regression coefficients as approximate reversal or mistake rates. 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 rate of affirmative decisions in the data. Suppose that, absent the bias toward negative autocorrelation in decisions, the rate of affirmative decisions 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 $R \equiv (\beta_0 - a) \cdot P(Y_{t-1} = 0) + (a - (\beta_0 + \beta_1)) \cdot P(Y_{t-1} = 1)$. To simplify, substitute $\beta_0 = a(1-\beta_1)$, so that the previous equation simplifies to $R \equiv -2\beta_1 a(1-a)$, which is positive since $\beta_1 < 0$.

If the correct decision is known, we can also estimate the fraction of decisions that are mistakes caused by the negative autocorrelation in decisions. Consider the alternative simple regression $Y_t = \widetilde{\beta}_0 + \widetilde{\beta}_1 Y_{t-1} + \gamma Y_{t,true} + e_t$. Let $\tau \equiv E[Y_{t,true}]$ be the rate of affirmative decisions in the data if all decisions were correct. Let $\delta \equiv E[1\{Y = Y_{true}\}]$ be the accuracy rate in the data. Taking expectations, $P(Y=1) = \left(\widetilde{\beta}_0 + \gamma\tau\right) / \left(1 - \widetilde{\beta}_1\right)$. Let $\widetilde{a} \equiv \left(\widetilde{\beta}_0 + \gamma\tau\right) / \left(1 - \widetilde{\beta}_1\right)$ be the rate of affirmative decisions in the data. Suppose that, absent the bias toward negative autocorrelation in decisions, the rate of affirmative decisions would still equal $\widetilde{a}$. 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 $\left(\widetilde{\beta}_0 + \gamma\tau - \widetilde{a}\right)$. If the previous decision was an affirmative, then the current decision is not likely enough to be an affirmative by the amount $\left(\widetilde{a} - \left(\widetilde{\beta}_0 + \gamma\tau + \widetilde{\beta}_1\right)\right)$. Therefore, the fraction of decisions that are reversed due to the negative autocorrelation is $\widetilde{R} \equiv \left(\widetilde{\beta}_0 + \gamma\tau - \widetilde{a}\right) \cdot P(Y_{t-1} = 0) + \left(\widetilde{a} - \left(\widetilde{\beta}_0 + \gamma\tau + \widetilde{\beta}_1\right)\right) \cdot P(Y_{t-1} = 1)$. To simplify, substitute $\widetilde{\beta}_0 + \gamma\tau = \widetilde{a}\left(1 - \widetilde{\beta}_1\right)$, so that the previous equation simplifies to $\widetilde{R} \equiv -2\widetilde{\beta}_1 \widetilde{a}(1-\widetilde{a})$, which is positive since $\widetilde{\beta}_1 < 0$.

The fraction of decisions that are mistakes caused by the negative autocorrelation is approximately $M = \widetilde{R}(\delta_0) - \widetilde{R}(1-\delta_0)$, where $\delta_0 = \delta + M$ is the accuracy rate if there were no negative autocorrelation in decisions. The mistake rate is the sum of the fraction of decisions that would have been accurate but are reversed due to the negative autocorrelation in decisions minus the fraction of decisions that would have been inaccurate but are reversed due to the negative autocorrelation in decisions. Note that, in extreme situations where the decision-maker is wrong more than half the time (e.g. $\delta < 0.5$), reversals can increase accuracy. Solving yields a mistake rate of $M = \frac{(2\delta-1)\widetilde{R}}{1-2\widetilde{R}}$.

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).[20]

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:

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

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,

[20] In this simple model of the gambler's fallacy in decision-making, agents form priors based upon previous decisions. In a more general 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.

$\alpha N - y_{t-1}$ 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 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[21] 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 through 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

[21]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/

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
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 movement
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

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* .

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, 2015, Incentivizing Calculated Risk-Taking: Evidence from an Experiment with Commercial Bank Loan Officers, *Journal of Finance* 70, 537–575.

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".

Gilovich, Thomas, Robert Vallone, and Amos Tversky, 1985, The Hot Hand in Basketball: On the Misperception of Random Sequences, *Cognitive Psychology* 17, 295–314.

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).

Green, Brett S., and Jeffrey Zwiebel, 2015, The Hot-Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments? Evidence from Major League Baseball, Working paper, University of California, Berkely and Stanford Graduate School of Business.

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

Hartzmark, Samuel M., and Kelly Shue, 2015, A Tough Act to Follow: Contrast Effects in Financial Markets, *Working Paper* .

Miller, Joshua Benjamin, and Adam Sanjurjo, 2014, A Cold Shower for the Hot Hand Fallacy, IGIER Working Paper 518, Innocenzo Gasparini Institute for Economic Research.

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.

Simonsohn, Uri, 2006, New Yorkers Commute More Everywhere: Contrast Effects in the Field, *The Review of Economics and Statistics* 88, 1–9.

Simonsohn, Uri, and Francesca Gino, 2013, Daily Horizons Evidence of Narrow Bracketing in Judgment From 10 Years of MBA Admissions Interviews, *Psychological science* 0956797612459762.

Simonsohn, Uri, and George Loewenstein, 2006, Mistake #37: The Effect of Previously Encountered Prices on Current Housing Demand, *The Economic Journal* 116, 175–199.

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

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.

Table I
Asylum judges: summary statistics

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

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

Table II
Asylum judges: baseline results

	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)	
β_1: Lag grant - grant					-0.0549***
					(0.0148)
β_2: Lag deny - grant					-0.0367**
					(0.0171)
β_3: Lag grant - deny					-0.00804
					(0.0157)
p-value: $\beta_1 = \beta_2 = \beta_3$					0.0507
p-value: $\beta_1 = \beta_2$					0.290
p-value: $\beta_1 = \beta_3$					0.0214
p-value: $\beta_2 = \beta_3$					0.0503
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	150,357	80,733	36,389	23,990	10,652
R^2	0.374	0.207	0.223	0.228	0.269

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 below 0.2 or above 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. The sample is further restricted to observations in which the current, previous, and previous-previous cases share the same defensive status. To retain sample size, we keep the restriction that the current and previous case must occur on the same day, but allow the previous-previous case to occur on the previous day. *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.

Table III
Asylum judges: heterogeneity

	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	23,990	23,990	22,965	22,965
R^2	0.229	0.229	0.229	0.247

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 II, Column 3. Standard errors are clustered by judge. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table IV
Loan officers: summary statistics

	Full Sample		Flat Incentives		Strong Incentives		Strongest Incentives	
Loan officer x loan observations	9168		1332		6336		1470	
Loan officers	188		76		181		89	
Sessions (6 loans per session)	1528		222		1056		245	
	Mean	S.D./(S.E.)	Mean	S.D./(S.E.)	Mean	S.D./(S.E.)	Mean	S.D./(S.E.)
Fraction of loans performing	0.65		0.66		0.65		0.65	
Fraction loans approved	0.73		0.81		0.72		0.68	
Fraction decisions correct	0.64		0.66		0.64		0.64	
Fraction performing loans approved	0.78		0.86		0.77		0.75	
Fraction non-performing loans approved	0.62		0.72		0.61		0.55	
Tetrachoric correlation	0.29***	(0.017)	0.29***	(0.047)	0.28***	(0.020)	0.32***	(0.040)
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

This table presents summary statistics on the sample of loan officers obtained from Cole et al. (2015) that we use in our decision-making analysis. The tetrachoric correlation is the correlation between the loan officer approval decision in the experiment and the indicator for whether the loan is a performing loan. The loan rating represents the continuous quality score loan officers assigned to each loan file during the experiment. This loan rating ranges from 0 to 100 and has been scaled down to be between 0 and 1.

Table V
Loan officers: baseline results

	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	7,640	7,640	2,615	2,615
R^2	0.0257	0.0536	0.0247	0.0544

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.

Table VI
Loan officers: heterogeneity

	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	7,640	7,640	7,640	7,640
R^2	0.0256	0.0281	0.0260	0.0256

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 V. Standard errors are clustered by loan officer x incentive treatment. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table VII
Loan officers: reactions to streaks

	Approve Loan Dummy	
	(1)	(2)
β_1: Lag approve - approve	-0.0751***	-0.165***
	(0.0216)	(0.0329)
β_2: Lag approve - reject	-0.0691***	-0.0955***
	(0.0236)	(0.0347)
β_3: Lag reject - approve	-0.0322	-0.0832**
	(0.0225)	(0.0332)
p-value: $\beta_1 = \beta_2 = \beta_3$	0.0178	0.00448
p-value: $\beta_1 = \beta_2$	0.703	0.0134
p-value: $\beta_1 = \beta_3$	0.00493	0.00300
p-value: $\beta_2 = \beta_3$	0.0483	0.688
Sample	All	Moderates
N	6,112	2,092
R^2	0.0290	0.0322

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 V . Standard errors are clustered by loan officer x incentive treatment. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table VIII
Baseball umpires: summary statistics

Number of called pitches following a previous called pitch	1,536,807
Number of called pitches following a consecutive previous called pitch	898,741
Number of games	12,564
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

This table presents summary statistics for the sample of MLB umpire calls that we use in our decision-making analysis. The sample represents all called pitches by MLB umpires from all games during the 2008 to 2012 seasons, covering 3.5 million pitches in 12,564 games, from 127 different home plate umpires. We restrict the sample to called pitches following a previously called pitch in the same inning. We classify a pitch as ambiguous if the location of the pitch is within 1.5 inches of the boundary of the strike zone. We classify a pitch as obvious if 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.

Table IX
Baseball umpires: baseline results

Strike	Full Sample		Consecutive Pitches	
	(1)	(2)	(3)	(4)
Lag strike	-0.00924***		-0.0146***	
	(0.000591)		(0.000972)	
β_1: Lag strike - strike		-0.0133***		-0.0208***
		(0.00104)		(0.00269)
β_2: Lag ball - strike		-0.0100***		-0.0188***
		(0.000718)		(0.00157)
β_3: Lag strike - ball		-0.00276***		-0.00673***
		(0.000646)		(0.00155)
p-value: $\beta_1 = \beta_2 = \beta_3$		1.49e-31		5.17e-22
p-value: $\beta_1 = \beta_2$		0.000423		0.414
p-value: $\beta_1 = \beta_3$		4.71e-25		3.07e-08
p-value: $\beta_2 = \beta_3$		3.79e-24		1.62e-21
Pitch location	Yes	Yes	Yes	Yes
Pitch trajectory	Yes	Yes	Yes	Yes
Game conditions	Yes	Yes	Yes	Yes
N	1,536,807	1,331,399	898,741	428,005
R^2	0.669	0.668	0.665	0.669

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.

Table X
Baseball umpires: ambiguous vs. obvious calls

Strike	Current Pitch Ambiguous		Current Pitch Obvious	
	(1)	(2)	(3)	(4)
Lag strike	-0.0347***		-0.00226***	
	(0.00378)		(0.000415)	
β_1: Lag strike - strike		-0.0479***		-0.00515***
		(0.0113)		(0.00101)
β_2: Lag ball - strike		-0.0324***		-0.00442***
		(0.00566)		(0.000773)
β_3: Lag strike - ball		-0.000838		-0.00283***
		(0.00563)		(0.000841)
p-value: $\beta_1 = \beta_2 = \beta_3$		1.74e-11		0.00573
p-value: $\beta_1 = \beta_2$		0.148		0.395
p-value: $\beta_1 = \beta_3$		0.0000205		0.0104
p-value: $\beta_2 = \beta_3$		5.02e-11		0.00507
Pitch location	Yes	Yes	Yes	Yes
Pitch trajectory	Yes	Yes	Yes	Yes
Game conditions	Yes	Yes	Yes	Yes
N	151,501	73,820	335,318	153,996
R^2	0.317	0.316	0.891	0.896

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 IX. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table XI
Baseball umpires: heterogeneity

	(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	898,741	898,154	894,779
R^2	0.665	0.665	0.665

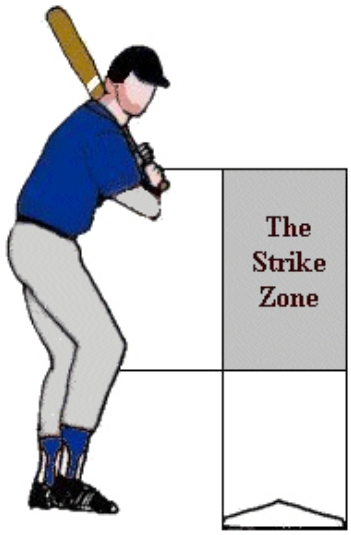
This table tests how our results differ depending on game conditions or umpire characteristics. The sample is restricted to consecutive called pitches. *Leverage* and *umpire accuracy* are represented as z-scores. *Leverage* is a measure developed by Tom Tango of how important a particular situation 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 IX. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table XII
Baseball umpires: treating teams "fairly"

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			-0.00821***
			(0.00188)
p-value: equality	6.70e-22	0.00158	0.0000736
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

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. p-values report tests for the equality of the reported coefficients. All control variables are as described in Table IX. Standard errors are clustered by game. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Figure I
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."