DRAFT – PLEASE DO NOT QUOTE OR CITE WITHOUT PERMISSION FROM THE AUTHORS.

Predicting Outcomes in a Sequence of Binary Events:
Belief Updating and Gambler's Fallacy Reasoning

Kariyushi Rao and Reid Hastie
Department of Behavioral Science, The University of Chicago Booth School of Business

April 20, 2020

Kariyushi Rao https://orcid.org/0000-0002-5027-8233

Correspondence should be addressed to Kariyushi Rao, Department of Behavioral Science, Chicago Booth School of Business, 5807 S Woodlawn Avenue, Chicago IL 60637. Email: kariyushi.rao@chicagobooth.edu.

Acknowledgements: This research was supported by funds from the Chicago Booth Graduate School of Business. The authors would like to thank the Chicago Booth community of scholars and many colleagues, most notably Maya Bar-Hillel, Diag Davenport, Samuel Hirshman, Emir Kamenica, Joshua Klayman, Andrew Meyer, Joshua B. Miller, Richard Thaler, Oleg Urminsky, Chen Xia, Kazuo Yamaguchi, and Nick Epley. We also wish to recognize our research assistants for their help categorizing our qualitative data: Helena Karas, Nicholas O'Donnell, and Leah Umanskiy. The first author is greatly indebted to Chris Wickens and Philipp Chapkovski for their timely and thoughtful guidance on programming experiments in oTree.

Abstract

We report on six experiments studying participants' predictions of the next outcome in a sequence of binary events. Participants faced one of three mechanisms generating 18 sequences of 8 events: a random mechanical bingo cage, an intentional goal-directed actor, and a financial market. We systematically manipulated participants' beliefs about the base rate probabilities at which different types of outcomes were generated by each mechanism. Participants either faced unknown (ambiguous) base rates, a specified distribution of three equiprobable base rates, or a precise, stationary base rate. Six target sequences ended in streaks of between two and seven identical outcomes. We focused on participants' predictions of the ninth, unobserved outcome in each of these target sequences. Across all generating mechanisms and prior belief conditions, the most common prediction pattern was best described as close-to-rational belief updating, producing an increasingly strong bias toward repetition of streaks. The exception to this generalization was for sequences generated by a random mechanical bingo cage with a precise, stationary base rate of .50. Under these conditions, participants exhibited a bias toward reversal of streaks. This effect was irrational, given our instructions on the nature of the generator. We conclude that the dominant judgment habit when predicting outcomes of sequences of binary events is reasonable belief updating. We review alternate accounts for the anomalous judgments of sequences produced by random mechanical devices with a precise, stationary base rate.

Keywords: binary sequences, streaks, hot hand, gambler's fallacy, belief updating

Predicting Outcomes in a Sequence of Binary Events:

Belief Updating and Gambler's Fallacy Reasoning

People spend a lot of time trying to predict the future. Perhaps the purest form of these inductive projections occurs when a person tries to forecast the next outcome following a sequence of similar binary events: heads or tails on the next coin flip, success or failure of a basketball player's next shot, or whether a company's stock price will rise or fall tomorrow. In these situations, the simplest strategy people follow is to predict "more of the same" (Soetens, Boeur, & Hueting, 1985). A somewhat more sophisticated strategy involves forming an impression of statistical relationships among past outcomes in order to predict more complex past-to-future patterns (Restle, 1966). But, the highest levels of reasoning involve inducing a causal explanation for the pattern in past outcomes, and relying on this causal mental model of the outcome-generating process to make forecasts for the future (Estes, 1964; Oskarsson, Boven, McClelland, & Hastie, 2009, for a review). The focus of the present research is on how these causal mental models influence forecasts of future outcomes in sequences of binary events.

For example, a gambler forms a mental model of roulette based on her past experience playing the game. This experience leads her to develop prior beliefs about the random nature of the outcomes produced by the wheel, and about the base rate probabilities at which the ball lands in a red or black pocket. A sports fan forms a mental model of basketball players based on his past experience watching his favorite team play. This experience leads him to develop a belief that athletes, as intentional, goal-directed actors, have control over their own performance outcomes. These abstract mental models are evoked when the gambler tries to predict what pocket the ball will land in next, or when the fan tries to predict whether his favorite player will make or miss his next shot.

We are interested in characteristics of the mental models people create for three different classes of outcome-generating processes (generators): random mechanical devices, intentional (human) actors, and social (financial) market processes. These three classes of generators have drawn the most attention from researchers investigating two different judgment patterns in people's predictions for the next outcome following a sequence of binary events. The first pattern occurs when people *increase* their expectation that a certain type of outcome will occur after observing that outcome repeat several times in a row. For example, the sports fan might show up to a Lakers game with a mental model of his favorite player, LeBron James, that includes prior beliefs about LeBron's base rate of success for field goal shots (about 50% at the time of writing). After watching LeBron successfully hit 5 field goals in a row, the fan *increases* his expectation that LeBron will succeed on his next (6th) field goal attempt *above* LeBron's base rate for landing field goals. This judgment pattern is often called the *hot hand*, after a belief among sports fans that players sometimes enter a "hot" state where their rate of success increases above their career (or season) average. The hot hand judgment pattern is most often observed in people's predictions for intentional actors (for reviews see Alter & Oppenheimer, 2006; Bar-Eli, Avugos, & Raab, 2006).

The second pattern occurs when people *decrease* their expectation that a certain type of outcome will occur after observing that outcome repeat several times in a row (Feller, 1968, p. 86ff). For example, the gambler arrives at a roulette table with a mental model of the game that includes prior beliefs about the base rate at which the ball lands in a red-colored pocket (about 47% in American-style roulette). After watching the ball land on red 5 times in a row, the gambler *decreases* her expectation that the ball will land in a red pocket on the next (6th) spin to some rate *below* the base rate at which the ball lands on red. This judgment pattern is called the

gambler's fallacy, because it is frequently observed in casino gambling situations (Croson & Sundali, 2005; Laplace, 1902/1814). The gambler's fallacy judgment pattern reliably occurs for a subset of observers making predictions for future events produced by random mechanical devices (for a recent review, see Reimers, Donkin, & Le Pelley, 2018). Interestingly, *both* the hot hand *and* gambler's fallacy judgment patterns are frequently observed in people's predictions for financial market processes (Conrad & Kaul, 1998; Forbes, 1995; Johnson, Tellis, & Macinnis, 2005).

Researchers have offered a variety of theoretical accounts to explain the hot hand and gambler's fallacy judgment patterns. Only a few of these accounts offer a unified framework that explains *both* the hot hand *and* the gambler's fallacy patterns. And, even fewer explain differences in prediction patterns *across classes of generators*. (Why is it that the gambler's fallacy pattern occurs most often when people make predictions for random mechanical devices, but the hot hand pattern occurs most often when people make predictions for intentional actors? And, how do we know which of these patterns will emerge when people are making predictions for financial market processes?) We believe one reason that a comprehensive account has not emerged is that researchers have focused on the wrong characteristics of people's mental models.

Prior experimental studies comparing people's predictions for different types of generators have focused on the qualitative descriptions of those generators (e.g. as a random mechanical device, or as an intentional, goal-directed actor). This work has often failed to consistently represent information about the base rates at which these different types of generators produce outcomes. Generators described as having stationary, well-known base rates (e.g. coins, dice, and roulette wheels), are often compared to generators whose base rates are described as uncertain or ambiguous (e.g. car salesmen, magicians, and basketball players). The

present studies are the first to systematically disentangle qualitative descriptions of the generator from prior beliefs about the generator's base rate. Our goal is to understand how these two sources of information influence people's predictions for future events. The question we are most interested in is whether it is necessary to go beyond reasonable updates of an estimated base rate in order to capture people's cognitions when predicting outcomes in sequences of binary events. To anticipate our conclusions, we will discover that simple updating of estimated base rates is the primary cognitive process producing increasing expectations of repetition (hot hand) as streak length increases for *all types of generators* when the base rate is uncertain or ambiguous. But, when the base rate is certain (stationary and well-known), many observers exhibit decreasing expectations of repetition (gambler's fallacy) specifically for random mechanical devices.

This article is organized as follows. In Section I, we review prior research on people's predictions for future outcomes in sequences of binary events. In Section II, we provide an overview of the present studies, as well as a description of common design elements. Section III presents the results of Studies 1A and 1B. Section IV presents the results of Studies 2A and 2B. Section V presents the results of Studies 3A and 3B. Section VI concludes by discussing the implications and limitations of our six studies, and identifying potential areas for future research.

Section I: Theoretical and Empirical Context for the Present Studies

The fact that a person can exhibit opposite judgment patterns after observing identical patterns of outcomes is a fascinating characteristic of people's behavior when forecasting events in binary sequences. The primary challenge for any theoretical analysis of this phenomenon is to provide a valid explanation for the differences in predictions following identical patterns produced by different generators. A second challenge is to account for differences when people

hold strong prior beliefs that the generator produces outcomes at a stationary, known rate, versus when people's prior beliefs about the generator's base rate are uncertain or ambiguous.

One class of explanations has made the most progress toward meeting both of these challenges. These explanations focus on observers' beliefs about the causal process that generates outcomes. Some of these accounts characterize the observer's beliefs as biased, reflective of some flawed reasoning process. Other accounts characterize the observer's beliefs as the result of a reasonable belief-updating process that reflects the true statistical properties of the observer's environment. We'll contrast these characterizations in the context of relevant empirical evidence. Then, we'll consider some methodological challenges in the extant experimental literature, and discuss the contribution of the present studies.

Mental Models of The Generators: Random Devices, Intentional Actors, and Markets

Random Mechanical Devices. People seem to hold incorrect beliefs about the causal operations of random mechanical devices. Nickerson (2002; also see Lecoutre, 1992) suggests that a common assumption people make about random generators is that they produce each outcome with equal probability (Blinder & Oppenheimer, 2008, provide experimental evidence supporting this suggestion). People also expect random generators to produce sequences with a high proportion of reversals, or alternation rates of about .60 (Ayton, Hunt, & Wright, 1989; Bar-Hillel & Wagenaar, 1991; Falk, 1981; Rapoport & Budescu, 1997; Reimers, Donkin, & Le Pelley, 2018).[1] When asked to judge sequences produced by a generator whose outcomes have unequal base rates, or that exhibit an alternation rate lower than .60, people tend to view those sequences as too streaky, and judge the outcome-generating process to be non-random (Gronchi

[1] The alternation rate is defined as $p(A) = (r - 1) / (n - 1)$, where r is the number of "runs" (a streak of identical outcomes), and n is the number of signals (outcomes) in the sequence. For example, the sequence *aabaaab* has $r = 4$ runs (*aa, bb, aaa,* and *b*), and $n = 7$ signals. The alternation rate for this sequence is $p(A) = (4 - 1) / (7 - 1) = 3/6 = .50$.

& Sloman, 2008; Lopes & Oden, 1987; Olivola & Oppenheimer, 2008; Scholl & Greifeneder, 2011). When asked to predict future outcomes for sequences produced by random generators, people expect the proportion of outcomes to reflect the population rate for each outcome type (i.e. to "balance out") even in short sequences, which results in a gambler's fallacy pattern of increasing expectations that streaks of identical outcomes will reverse (cf. Boynton, 2003; Gronchi & Sloman, 2008).

Intentional Actors. People tend to expect performances by skilled human actors to exhibit "streaky" patterns of repeated outcomes. When asked to predict future outcomes for sequences produced by intentional actors, people exhibit a hot hand pattern of increasing expectations that streaks of identical outcomes will repeat (Alter & Oppenheimer, 2006; Bar-Eli, Avugos, & Raab, 2006; cf. Boynton, 2003; Gilovich, Vallone, & Tversky, 1985; Fischer & Savranevski, 2015; Vergin, 2000). Some researchers suggest people believe there is something "special" about *human* performance (Ayton & Fischer, 2004), or about the *intentional mind* of human actors (Caruso, Waytz, & Epley, 2010; Roney & Trick, 2009). But, people only seem to exhibit hot hand beliefs for human actors whose performances they specifically perceive as *non*-random (Burns & Corpus, 2004; Tyszka, Zielonka, Dacey, & Sawicki, 2008).[2]

[2] Burns and Corpus (2004) and Tyska, Zielonka, Dacey, and Sawicki (2009) both asked experimental participants to judge sequences produced by two different human actors: one rated as *more* random (your little sister shooting baskets in Burns & Corpus; a fortune-teller in Tyska et al.), and another rated as *less* random (a competitive car salesman in Burns & Corpus; a basketball player shooting baskets in Tyska et al.). Participants exhibited hot hand beliefs in their predictions for the *less* random human actors. But, the judgment pattern for the *more* random human actors was either neutral (Burns & Corpus) or inconsistent (Tyska et al.). Importantly, the little sister shooting baskets in Burns and Corpus's study was explicitly described as *intending* to improve her success rate, and the Polish participants in Tyska and colleagues' experiment were familiar with effortful performances by fortune-tellers who use various physical props to underscore their particular skill in accurately forecasting the future. So, the differences in participants' predictions for these generators cannot be explained by participants failing to perceive the little sister or fortune-teller as intentional, goal-directed actors.

Markets. People apparently hold a mixed bag of beliefs about markets. There is evidence that both novices (De Bondt, 1993) and experts (Baquero & Verbeek, 2015; Barberis, Shleifer, & Vishny, 1998; Shanthikumar, 2012) expect streaks of market outcomes (e.g. individual stock price movements) to repeat. There is also evidence that both novices (Anderson & Sunder, 1995) and experts (De Bondt, 1991; De Bondt & Thaler, 1990; Durham, Hertzel, & Martin, 2005; Loh & Warachka, 2012) expect streaks of market outcomes to reverse. People's predictions for streaks of market outcomes might also be influenced by the alternation rate of prior outcomes. Bloomfield and Hales (2002) present experimental evidence that people expect streaks preceded by a sequence with low alternation to repeat, while those preceded by a sequence with high alternation are expected to reverse. Novice and expert investors also seem to have different mental models for small companies versus large, and for young companies versus old (Bulkley & Harris, 1997; Burns, 2003).

Theoretical Accounts of Prediction Behavior

Heuristics and Biases. The heuristics and biases research program produced two related concepts that are often evoked to explain people's judgment behavior when forecasting outcomes in sequences of binary events. The *representativeness heuristic* describes an observer's tendency to judge the likelihood of an outcome based on how well that outcome represents "evidence" about the specific event the observer is trying to predict (Kahneman & Tversky, 1973).[3] The

[3] We interpret the authors' use of the term "evidence" to mean individuating information about the specific event an observer is asked to predict. In Study 1, the authors ask participants to predict the area of graduate study chosen by a student, Tom, described in a vignette. In this case, the specific event is the selection of an area of study by Tom, and the "generator" here is something like an abstract model of a graduate student producing choices of study. The vignette presents as "evidence" a description of Tom containing many stereotypic characteristics of computer science and engineering majors. This is the individuating information about the event "Tom chooses an area of study." In predicting Tom's area of study, participants ranked computer science and engineering (fields with a relatively low base rate of selection by graduate students) higher than humanities and business (fields with a relatively high base rate of selection by graduate students). The authors argue that the stereotypic information about Tom was non-diagnostic, and, therefore, participants should have ignored it and relied on the base rate of selection into each field of study.

Law of Small Numbers describes an observer's expectation that small samples of outcomes will reflect the (statistical) parameters of the population from which they were drawn (Tversky & Kahneman, 1971). For example, people expect a small sample of coin flips to have an equal number of Heads and Tails outcomes, because they believe that the rate at which Heads and Tails are produced in the population of fair coin flips is p(Heads) = p(Tails) = .50. An observer asked to predict the next outcome following a sequence containing a *dis*proportionate number of Heads outcomes (e.g. HTHHH) will assign higher probability to Tails, because that would make the resulting sequence (HTHHHT) more *representative* of the .50 rate she holds in her mental model of random devices like coins.

The representativeness and Law of Small Numbers heuristics only provide accounts for the gambler's fallacy pattern in people's predictions. Some extension is required to also account for the hot hand pattern in predictions for nonrandom generators, like intentional (human) actors. Gilovich, Valone, and Tversky (1985) proposed that when observers see a sequence that does *not* exhibit a high rate of alternation, they decide the generator (e.g., a professional athlete) must be a non-random mechanism, and shift to naïve expectations that outcomes will repeat at a high rate. We have several reservations about this interpretation, including the odd idea that a sports fan would have a default expectation that an athlete would behave like a random device, as well as problems with the data analysis the authors cite as support for this interpretation (recently noted by Miller & Sanjurjo, 2018b).

Rabin (2002; Rabin & Vayanos, 2010) offers an alternative extension of the Law of Small Numbers that makes more sense to us. He proposed that the observer assumes outcomes are sampled *without replacement* (Estes, 1964; Fiorina, 1971; Morrison & Ordeshook, 1975; and Restle, 1961, also proposed *sampling without replacement* models). In Rabin's account, rational

belief updating is the central cognitive process underlying predictions of events in binary sequences, but the observer has a non-standard mental model of the generator's causal process. The observer begins with correct prior beliefs about the distribution of possible base rates and is rationally Bayesian. But, instead of assuming that outcomes are generated by an abstract *independent* and identically distributed random process, she imagines they are drawn *without replacement* from a small urn containing N signals, $s_i \in \{a, b\}$, in proportion to the generator's base rate, θ. This means that the observer expects the urn to contain exactly θN *a* signals, and $(1 - \theta)N$ *b* signals. When the observer sees a short sequence of signals from a generator with a *known*, *stationary* rate, she reasons about them in terms of an urn whose contents are depleted as the sample of signals is drawn. Following repeated draws of one signal type, the observer believes there are *fewer* of that signal type left in the urn, and therefore assigns increasingly lower probability to subsequent draws of that type, producing a gambler's fallacy judgment pattern.

In the case of a generator with a known, stationary rate, the observer does not find it *a priori* plausible that the proportion of signals in the urn might change over time. However, when the observer is confronted with a generator having an *unknown* or *ambiguous* rate, she might indeed find such a change plausible. In this situation, the observer defaults to the expectation that the generator is like a random mechanical device with a base rate of .50. But, if she encounters a streak of identical signals, she overreacts to what she perceives as too few reversals. She then adopts an *additional* belief that the generator's rate changes over time (Rabin & Vayanos, 2010, discussion starting on p. 746). Conditional on the sequence of signals she has just observed, she updates her beliefs to reflect the most likely base rate to have produced that

sequence, again in a Bayesian manner.[4] As a given signal type continues to repeat, increasing the length of the streak, the observer continues to update her beliefs, assigning an even higher rate to the signal type repeated in that streak. Thus, as streak length increases, the observer's updated predictions will exhibit a hot hand pattern.

This hypothesized shift from initially expecting reversal to eventually expecting repetition is unique to Rabin's Small Urn Model. A few empirical reports provide suggestive evidence that people's predictions for binary sequences do exhibit the sort of U-shaped pattern produced by such a shift (especially for long streaks of more than 10 binary outcomes; Altmann & Burns, 2005; Asparouhova, Hertzel, & Lemmon, 2009; Edwards, 1961; Lindman & Edwards, 1961; Jarvik, 1951; Nicks 1959; Rao, 2009; Suetens, Galbo-Jorgensen, & Tyran, 2016; Tyszka, Markiewicz, Kubińska, Gawryluk, & Zielonka, 2017; and see early laboratory studies summarized in Lee, 1971, pp. 163-167). With reference to the apparent prevalence of hot hand beliefs among observers judging sequences produced by intentional actor, Rabin suggests that people are simply *less certain* about the generator's causal process. In other words, people have weak prior beliefs about the base rate at which an intentional actor produces different types of outcomes, and therefore people engage in belief updating about the base rate early in their prediction strategy.

With Rabin's extension, the representativeness heuristic and the Law of Small Numbers seem to provide a neat explanation of both the hot hand and gambler's fallacy judgment patterns. People assign higher probability to reversal of streaks produced by random mechanical devices, because they think that sort of pattern is more representative of their mental model of random mechanical devices. People assign higher probability to repetition of streaks in human

[4] Interested readers may find a discussion of the belief-updating process that leads the observer to overinfer the likely extremity of the base rate in Section IV of Rabin (2002, p. 788).

performance, because they think that sort of pattern is more "representative" of their mental model of intentional human actors. People exhibit mixed judgment patterns in their predictions for markets, because even experts can't agree about the presence or absence of patterns in market data. These explanations are very close to *post hoc* descriptions of the patterns researchers observe in their experimental and observational data ("People expect generators of type X to produce pattern Y, so that must mean pattern Y is part of their mental model for type X generators."). However, the representativeness heuristic and the Law of Small Numbers can't specify people's mental models *a priori*, or tell us much about their origin (cognitive, environmental, or otherwise).[5] These heuristics also don't provide clear predictions when people are faced with a new type of generator for which we have not previously accumulated judgment data (or in cases where contradictory patterns are observed in those data, as is the case in people's predictions for markets).

Experience and Education. One explanation for where the different expectations come from is an ecological account that posits people's beliefs about different generators result from a learning process that reflects the true statistical properties of their environment. Hahn & Warren (2009) present a version of this account (related explanations are also presented by Kareev, 2000; Miller & Sanjurjo, 2018a; Reimers, Donkin, & Le Pelley, 2018; and Sun & Wang, 2010). The authors remind us that people don't experience infinite sequences of outcomes (inside *or* outside of the laboratory). People have limited attention and finite experience. While a theoretical i.i.d. Bernoulli process will produce an *infinite* sequence of outcomes in which all exact orderings of substrings (e.g. HHHH, HHTT) are equally likely, the same is not true for

[5] Though Rabin presents a coherent theoretical account of how gambler's fallacy and hot hand patterns arise, he doesn't provide specific guidance for how we might predict whether or not an observer will find it *a priori* plausible that a generator's rate changes over time. For that guidance, we again are referred back to the judgment data that's been previously accumulated for different types of generators.

finite samples of sequences. Given finite samples, the probability of observing a given substring depends on the number of different realizations of the sample in which that substring occurs. To put it another way, imagine an observer watching sequential tosses of a fair coin. It will take longer (about 30 tosses) for this observer to encounter the substring HHHH than to encounter HHTT (about 16 tosses). If it is known *ex ante* that the observer will only watch 20 tosses total, then it is more likely the observer will encounter HHTT than HHHH.

Reversal expectations for random generators probably also result from scholastic training in mathematics classrooms. The pedagogic methods most often used to teach concepts related to probability and randomness reinforce reliance on the representativeness heuristic when judging sequences produced by random mechanical devices (Amir & Williams, 1999; Batanero, Chernoff, Engel, Lee, & Sánchez, 2016; Borovcnik and Peard, 1996; Harradine, Batanero, & Rossman, 2011; Hawkins & Kapadia, 1984; Jones, 2004; Konold, 1995; Meletiou-Mavrotheris, 2007; Morsanyi, Handley, & Serpell, 2013; Shaughnessy, 1992; Shaughnessy, Canada, & Ciancetta, 2003; Steinbring, 1990).[6]

Whether everyday experience or formal education leads people to build an assumption of negative serial correlation into their mental models of random generators, it doesn't seem like reversal predictions are automatic. The most common prediction pattern for very young children is to expect repetition of streaks (hot hand). The gambler's fallacy starts to show up in elementary school, and peaks among college students (Bogartz, 1965; Chiesi & Primi, 2009; Derks & Paclisanu, 1967; Estes, 1962; Craig & Meyers, 1963). People also take longer to predict reversal than repetition, and are more likely to predict repetition than reversal under time

[6] Some readers might find this statement confusing or controversial. We'll return to this point, and provide additional context, at the end of the present article when we reflect on the results of our experiments, particularly Studies 2A and 2B.

constraint or cognitive load (Braga, Ferreira, Sherman, Mata, Jacinto, & Ferreira, 2018; Diener & Thompson, 1985; Militana, Wolfson, & Cleaveland, 2010; Tyszka, Markiewicz, Kubińska, Gawryluk, & Zielonka, 2017).

People's mental models of intentional actors actually seem to reflect the true behavior of these generators pretty well. Gilovich, Vallone, and Tversky's (GVT, 1985) seminal paper generated a lot of buzz for apparently demonstrating that basketball fans' belief in the hot hand was irrational. A flurry of empirical studies followed, some confirming GVT's results by demonstrating that serial correlation in sequences of human performance did not exceed chance levels, and others contradicting the original results by demonstrating consistent, significant positive recency in sequences of skilled human performance (for reviews see Alter & Oppenheimer, 2006; Bar-Hillel, Avugos, & Raab, 2006). Recently, Miller and Sanjurjo (2018b) identified a significant error in GVT's original analysis of basketball shooting data, as well as in several replications of GVT's results. Miller and Sanjurjo's re-analysis of these studies revealed, "significant evidence of streak shooting, with large effect sizes," (*ibid*, p. 2022) in GVT's basketball data, as well as "hot hand effect sizes [that] are consistently moderate to large," (*ibid*) in the data from replications by Avugos, Bar-Eli, Ritov, and Sher (2013a) and by Kohler and Conley (2003).

It's difficult to come up with some sort of "ground truth" for the behavior of financial market generators. There's evidence of both positive and negative serial autocorrelation in stock market outcomes (Conrad & Kaul, 1998; De Bondt & Thaler, 1989; Jegadeesh & Titman, 2011). Both momentum (betting on repetition) and contrarian (betting on reversal) investment strategies yield statistically significant profits in some market segments, and are equally likely to be

successful (Conrad & Kaul, 1998). So, it's hard to say that either hot hand or gambler's fallacy predictions for stock market outcomes are unreasonable.

We find the ecological accounts of people's prediction behavior compelling, but these accounts suffer from the same limitation as the heuristics and biases perspective: It's not clear *a priori* what "real world" events are sources of influence on current predictions for different sequences. For example, how can we anticipate people's prior beliefs about the field goal success rate of a basketball player? Should we focus on the success rate for field goal shots taken over the observer's lifetime of experience with basketball, or only those taken by the person's favorite team, or by her favorite player? Or might the observer be relying on experiences across multiple sports or goal-directed achievement endeavors? Without rules to predict what experiences people might draw upon to form their mental model of a given generator, it's difficult to derive hypotheses about their judgments of the sequences that generator produces.

Our favored hypothesis about when to expect observers will predict repetitions versus reversals is based on Rabin's Small Urn Model. Rabin proposed gambler's fallacy, reversal, predictions will arise when people believe the generator has a *known*, *stationary* base rate, but that hot hand, repetition predictions will arise when people are *uncertain* about the base rate of the generator (or when they believe that rate may change over time). We like this interpretation because it begins to explain differences in participants' prediction patterns for different types of generators. In the extant literature on hot hand and gambler's fallacy judgments, we've observed a frequent tendency for experimenters to provide base rate information that is inconsistent across generator types.

Methodological Challenges in The Extant Literature

Most experimental studies investigating predictions of binary events in sequences present slightly different information about random versus intentional generators. Instead of clearly specifying identical base rates for each generator, experimenters either rely on participants' prior beliefs (e.g. that a fair coin has a stationary .50 base rate, or that a basketball player performs at an unspecified, perhaps "typical," rate), or they provide base rate information that may be interpreted as *stationary* for random devices, but *shifting* for intentional actors. For example, Braga and colleagues (2018) compare predictions for coin flips to those for athletic performances. The authors provided explicit information that the coin was *fair* (stationary .50 base rate), and that flipping the coin was a random process. Their description of the athletes provided no information about the athletes' performance rates, and explicitly stated that these rates shift as the athlete ages. Burns and Corpus (2004) compare predictions for a roulette wheel to those for a competitive car salesman and a little sister shooting baskets. Participants were informed that each generator had produced each type of outcome on 50/100 trials. It was not clear whether this .50 rate was stationary or shifting, and both the car salesman and little sister were described as attempting to improve their performances over time.

If experimenters consistently present clear, concrete information for random mechanical devices, but not for intentional actors, then, intuitively, we ought to expect differences in participants' judgment patterns for these types of generators.[7] In each of the present studies, we provide participants with identical information about each generator's base rate. Across studies, we systematically manipulate participants' level of certainty about the generators' base rate. In this way, we provide a direct test of Rabin's conjecture that uncertainty about the base rate

[7] In experiments that elicit predictions for market outcomes, the description of the market generator sometimes provides explicit information about the base rate (e.g. by evoking the concept of a fair coin, or describing a random walk process), and sometimes provides no information about the base rate. These discrepancies are probably one source of variation in participants prediction habits for market generators.

determines whether people exhibit hot hand or gambler's fallacy patterns in their predictions for sequences of binary events.

Section II: Overview of Present Empirical Studies

In the present studies, participants are shown sequences of 8 binary outcomes, and are asked to predict the direction of the 9th (next) outcome in each sequence. This simple "What's next?" binary event prediction task is a popular experimental model for the study of cognitive processes that underlie important everyday judgments. It provides a simplified analogue to many everyday judgments, and participants easily grasp the constraints and objectives of this familiar task. There is little ambiguity in the instruction to predict the outcome of the next event in a sequence of similar events. And, the structure of the task makes it straightforward to study the way information from bottom-up data and top-down abstractions interact to produce a unitary response. There is also considerable variety in participants' prediction strategies, and this variety expresses itself in a form that facilitates clustering of participants (or even individual prediction trials) into meaningful blocks of similar strategies. We will show that this task also does a good job of separating adaptive, even rational strategies from response habits that appear to be maladaptive and irrational.

In each study, participants judge sequences produced by one of three different generators: 1) a bingo cage filled with red and blue balls (random mechanical device), 2) an investment portfolio analyst (intentional, goal-directed actor), and 3) a publicly traded company (market process). The bingo cage produces Red or Blue outcomes, and the investment analyst and company produce Up or Down outcomes. Across 6 target experimental sequences, we manipulate the length of a terminal streak of identical outcomes (e.g. Red-Red, Down-Down-Down). Each participant sees one sequence ending in each of the following streak lengths: 2, 3,

4, 5, 6, and 7.[8] Extrapolating from previous empirical results comparing predictions across generators, we expect participants assigned to judge sequences produced by the bingo cage will be more likely to expect reversals of these streaks, and participants assigned to judge the sequences produced by the investment analyst will be more likely to expect repetition (we do not have a clear prediction for participants assigned to judge sequences produced by the publicly traded company, due to the mixed results for social or market processes in the literature).

Participants are asked to predict the next (9th) outcome in each sequence either by making a dichotomous choice, or using a continuous probability scale. There is some evidence that people are more likely to engage in intuitive reasoning when asked to make a dichotomous choice, and more likely to engage in analytical reasoning when asked to make probability ratings on a numerical scale, so we use both response formats as a robustness check.[9] If gambler's fallacy and hot hand patterns emerge as a result of people's reliance on intuitive (heuristic) reasoning, we should see more extreme preferences for reversal or repetition among participants asked to respond with a dichotomous choice than among those asked to respond using a numerical scale.

Within each Study, we are careful to provide identical information about the base rates of the three generators.[10] In Study 1, we provide no information about any of the generators' base rates. We do not indicate the ratio of red to blue balls in the bingo cage, the rate at which the investment analyst's portfolio increases or decreases in value, or the rate at which the company's

[8] We use multiple versions of each target sequence, varying the type of outcome in the streak as well as the pattern of outcomes preceding the streak. Please see Appendix A: Stimuli and Appendix B: Procedure for a detailed description of the stimuli and the method of randomization used for presenting these stimuli to participants.
[9] Perhaps the strongest empirical support for this hypothesis comes from some surprising results reported by Murphy and Ross (2010) in which a vast majority of respondents shifted from an irrational selective attention strategy when making dichotomous responses to an almost rational Bayesian strategy when making numerical category-based probability inferences (also see Hammond, Hamm, & Grassia, 1987; Önkal & Muradoglu 1996; Windschitl & Wells, 1996).
[10] The experimental instructions provided to participants in each study can be found in Appendix B: Procedure.

stock price increases or decreases. Following the implications of Rabin's (2002) and Burns's (2002) analyses, we expect participants' expectation of repetition to increase with the length of the terminal streak at the end of each target experimental sequence. In Study 2, we provide a stationary base rate of .50 for all three generators. Relying again on Rabin's (2002) model, we hypothesize that participants will initially *decrease* their expectation of repetition across the target experimental sequences with shorter terminal streaks, and then eventually *increase* their expectation of repetition as streak length increases. In Study 3, we specify the same distribution of possible base rates for all three generators. Given specific information about the distribution of possible rates, we expect participants to follow a Bayesian pattern of belief-updating, increasing their expectation of repetition as streak length increases.

Participant Recruitment

Participants in the present studies were sampled from Amazon Mechanical Turk (MTurk). Participants were required to live in the United States, and have a Human Intelligence Task (HIT) approval rate of at least 95% over at least 5 previously completed HITs. Participants were randomly assigned to one of three Conditions: Bingo, Analyst, or Stock. No participants who completed the full experimental procedure were excluded from the analyses in any of our experiments. No participant took part more than once in each of our experiments, or in more than one of our experiments.

Procedure

The experimental procedure was implemented using the oTree platform (Chen, Schonger, & Wickens, 2016). Participants first read the instructions for the procedure, and then were required to answer 4-5 comprehension questions correctly before they were allowed to begin the

experimental session.[11] During the experimental procedure, participants were shown 18 sequences of 8 outcomes, one sequence on each of 18 trials. The 6 target experimental sequences ending in a streak were mixed with 12 filler sequences ending in a reversal (e.g. Red-Blue, Down-Up). Filler sequences were used to balance out the frequency of streaks and proportion of signal types across rounds of the experiment.[12] On each trial, participants were instructed that they were observing a "new" sequence of 8 consecutive outcomes, as opposed to a continuation of the 8 outcomes observed in the previous trial.[13] Each outcome was then revealed one at a time, starting from the left-hand side of the screen, with a one-second delay between the appearance of each outcome.[14] Participants were asked to wait until all 8 outcomes were revealed, and then to predict the next (9th) outcome, either by making a dichotomous choice, or using a continuous probability scale.[15] No feedback was provided after the participants made each prediction (the identity of the 9th outcome was not revealed). After completing the experimental procedure, participants answered 5 questions testing their knowledge of basic probability and their financial literacy as well as a brief demographics questionnaire (age,

[11] If participants answered any questions incorrectly, they were asked to check their answers and correct any mistakes. Participants were allowed to attempt the questions as many times as they wished. This process ensured that no participants started the experimental session without first answering all of the comprehension check questions correctly. A detailed description of the instructions, as well as the comprehension check questions, can be found in the Appendix B: Procedure.

[12] Filler sequences were randomly selected from a pool of 24 sequences ending in reversal. Details can be found in Appendix A: Stimuli. On the first trial, participants always saw one of the filler sequences. The 11 remaining filler sequences, and 6 target experimental sequences, were then presented in random order on trials 2 through 18.

[13] On each trial, participants in the Bingo Condition see the instruction, "Here is a new round of draws by the mechanical bingo machine," and participants in the Analyst and Stock Conditions see the instruction, "Here is a new [analyst / company]."

[14] Barron and Leider (2010) found that gambler's fallacy patterns of prediction emerged when sequences were presented sequentially, but not when they were presented all-at-once, so it is important to note that we are using sequential revelation in all of our experiments.

[15] The specific question prompts, as well as screenshots of the experimental interface, can be found in Appendix B: Procedure.

gender, and highest degree).[16] Participants were also asked to provide a qualitative description of the strategy they used to make their predictions.[17]

Section III: Studies 1A and 1B

Our goal in Studies 1A and 1B was to gather basic data about how people make judgments given verbal descriptions of our three generators (bingo cage, investment analyst, and public company) without any information about the generators' base rates.

Participants

144 participants (M_{AGE} = 36.02, SD_{AGE} = 11.23, N_{FEMALE} = 58) were recruited for Study 1A. Participants took 18.69 minutes on average (SD = 10.02) to complete Study 1A. 300 participants (M_{AGE} = 35.43, SD_{AGE} = 11.79, N_{FEMALE} = 143) were recruited for Study 1B. Participants took 17.75 minutes on average (SD = 9.65) to complete Study 1B. Participants in both studies were paid $2.50 upon approval of their completed tasks. (Due to the noisiness of binary choice data, our Study 1B participant sample is twice the size of that in Study 1A.[18])

Method

Participants in Study 1A made their predictions using a continuous numerical probability scale. Participants in Study 1B made their predictions as a dichotomous choice. Participants in both studies were randomly assigned to one of three experimental conditions, defined by the description and base rate of the generator. In the BingoUnknown Condition, the events in each sequence were described as draws (Red/Blue) made *with replacement* by a mechanical bingo

[16] Please see the *Individual Differences* section of the Online Supplement for a detailed description of the probability, financial literacy, and demographic questions. Interested readers may also find summary statistics and correlation matrices for these measures in the *Individual Differences* section. There were no significant relationships between any of these measures and the results of the present experiments, so they are not discussed further here.

[17] A summary of these qualitative responses can be found in Appendix C: Verbal Reports.

[18] For this and all following studies, our target recruitment numbers were set in advance based on rules of thumb (see, e.g., Cohen, 1988; Green, 1991; Harris, 1985), and subject to budgetary restrictions. We continued recruitment until we came as close as possible to our recruitment targets. We did not analyze any partial data.

machine from a cage containing 100 red and blue balls. No information was provided about the ratio of red to blue balls. In the AnalystUnknown Condition, the events in each sequence were described as quarterly changes (Up/Down) in the value of a particular investment analyst's portfolio. No information was provided about the rate at which the analyst's portfolio increased or decreased in value. In the StockUnknown Condition, the events in each sequence were described as quarterly changes (Up/Down) in a particular company's stock price. No information was provided about the rate at which the company's stock price increased or decreased. On each trial, participants were instructed that they were viewing a *new* sequence ("here is a new round of draws by the bingo machine," "here is a new [analyst / company]") to reinforce the idea that they were not seeing a continuation of the previous trial's sequence.

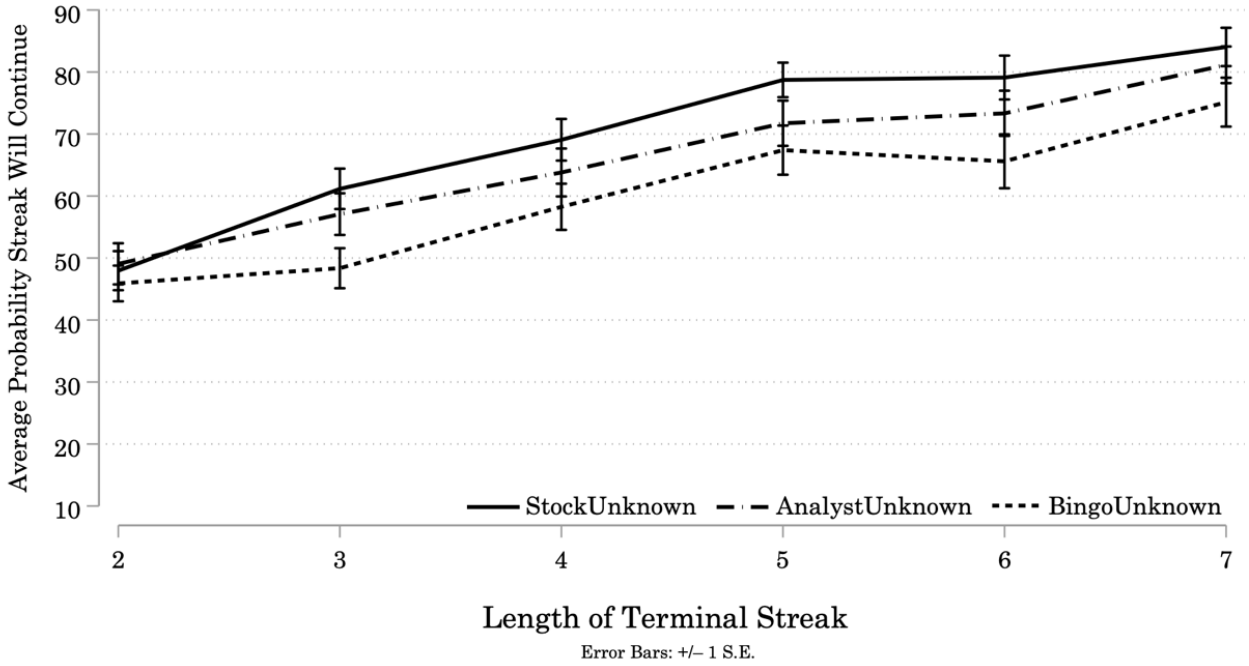
Results: Study 1A

Figure 1 shows the average probability participants assigned to *repetition* of the streak at the end of each target sequence. On average, participants in all three Conditions assigned greater than 50% probability to repetition of the terminal streaks. Participants' expectations of repetition also increased with Streak Length. We conducted a one-way mixed ANOVA to test the effects of one between-subjects variable (Condition), and one within-subjects variable (Streak Length) on participants' predictions about the probability that the terminal streak at the end of a given experimental sequence would repeat. The probability participants assigned to repetition of the terminal streak was the focus of analysis (average ratings in Table 1).[19]

[19] Participants' predictions over the filler sequences (all ending in a reversal) were *not* included in this or any subsequent analysis. Interested readers may visit the *Filler Sequences* section of the Online Supplement to find summary statistics for participants' predictions over the filler sequences.

Figure 1

Study 1A: Average Probability Participants Assigned to Repetition of Terminal Streaks, by Streak Length and Condition



Note. Results of Study 1A. Solid and patterned lines represent the average probability assigned by participants in each Condition to the event that the next (9th) outcome will repeat the streak of identical signals they observed at the end of each experimental sequence. Error bars represent +/– 1 standard error. Starting at Streak Length 3, participants in all three Conditions assigned greater than 50% probability to the event that the next signal would repeat the streak, and these probabilities increased with Streak Length. N = 144 (StockUnknown = 44; AnalystUnknown = 50; BingoUnknown = 50).

Table 1

Study 1A: Average Probability Participants Assigned to Repetition of Terminal Streaks, by Streak Length and Condition

		Average Probability Terminal Streak Will Repeat					
Condition	N	Streak = 2	Streak = 3	Streak = 4	Streak = 5	Streak = 6	Streak = 7
AnalystUnknown	50	49.06 (23.47)	57.08 (23.73)	63.80 (27.34)	71.74 (25.80)	73.34 (25.78)	81.16 (20.87)
StockUnknown	44	47.95 (20.90)	61.16 (21.60)	69.07 (22.28)	78.73 (18.48)	79.11 (23.49)	84.05 (20.51)
BingoUnknown	50	45.90 (20.34)	48.36 (22.80)	58.26 (26.35)	67.40 (28.05)	65.60 (30.73)	75.14 (27.89)
All	144	47.64 (21.53)	55.30 (23.25)	63.49 (25.74)	72.37 (24.93)	72.42 (17.34)	79.95 (23.58)

Note: Standard errors in parentheses.

There was a significant effect of Condition on participant predictions ($F(2, 141) = 4.11$, $p = 0.018$). Bonferroni-corrected pairwise comparisons revealed that predictions made by participants in the BingoUnknown Condition were significantly lower than predictions made by participants in the StockUnknown Condition (Mean Difference = –9.90, $p = 0.016$). However, predictions made by participants in the AnalystUnknown Condition were *not* significantly different than predictions made by participants in the BingoUnknown Condition (Mean Difference = 5.92, $p = 0.247$), *or* by participants in the StockUnknown Condition (Mean Difference = –3.98, $p = 0.770$).

There was a significant main effect of Streak Length on predictions, ($F(4.43, 623.98) = 58.46$, $p < 0.000$).[20] Longer Streak Lengths were assigned higher probability of continuing. This finding is consistent with the notion that participants are updating their estimates of the base rate given information in the event signals. It is also consistent with Rabin's (2002) model,

[20] Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(14) = 49.03$, $p < 0.000$); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.89$).

which predicts hot hand beliefs will arise over longer streaks of identical signals. There was no significant interaction between Streak Length and Condition ($F(8.85, 623.98) = 0.64$, $p = 0.758$). Participants faced with a generator described as an intentional actor updated their beliefs in a similar fashion to participants faced with a generator described as a random mechanical device or a market. When the base rate of the generator is ambiguous (unknown), intentionality is not a necessary condition for expectations of repetition to increase.

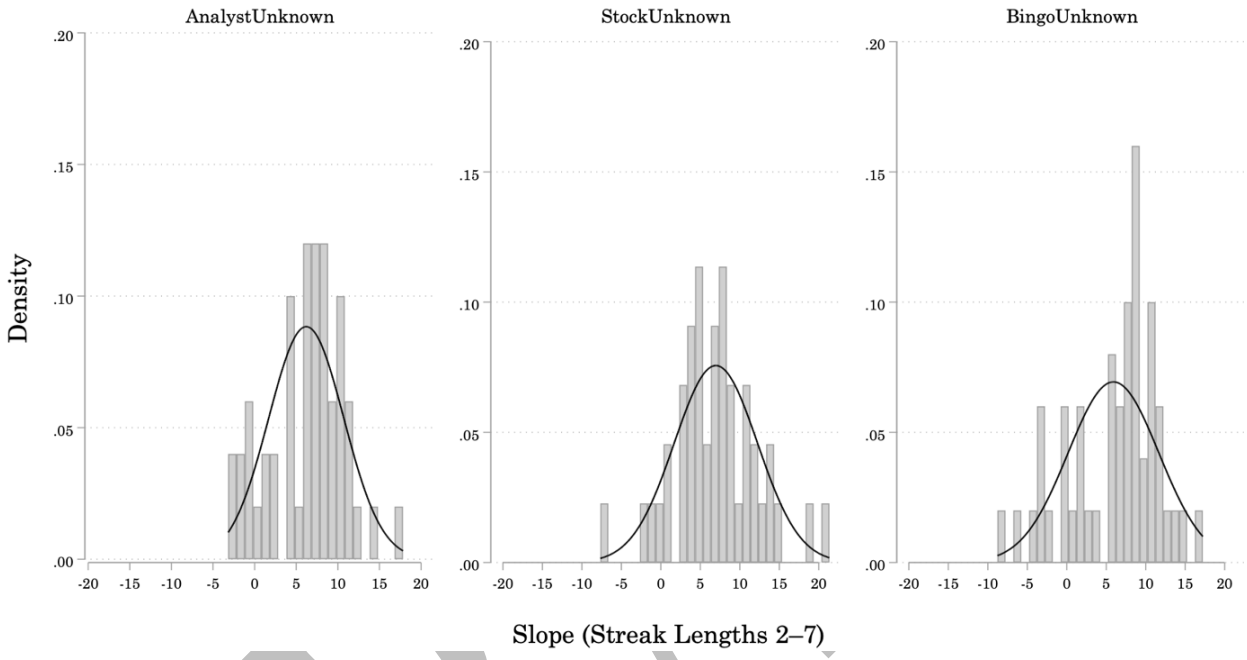
To get a sense of the differences between individual participants' prediction strategies we took a look at the slopes (coefficients) of linear regressions fitted to each participant's predictions over the target sequences ending in streaks of length 2 through 7. A positive slope indicates that the participant increases the probability assigned to repetition of the terminal streak as the length of that streak increases (hot hand pattern). A negative slope indicates the opposite strategy, that a participant decreases the probability assigned to repetition of the streak as the length of that streak increases (gambler's fallacy pattern). There was some heterogeneity in individual participants' prediction strategies, as measured by the slopes of their predictions.

Figure 2 presents the distribution of fitted slopes for individual participants' predictions over the target sequences. The distributions are centered above 5 in all three Conditions. On average, each incremental increase in the length of the terminal streak is related to an increase of about 5% in the probability participants assigned to repetition of that streak. The dominant strategy in all three Conditions is a positive slope, which we interpret as participants updating their beliefs about the base rate of the generators as Streak Length increased. A minority of participants exhibit negative slopes across their predictions, decreasing their expectations of repetition as Streak Length increased: 14% in the AnalystUnknown Condition, 7% in the

StockUnknown Condition, and 20% in the BingoUnknown Condition. This pattern of predictions is consistent with gambler's fallacy beliefs.

Figure 2

Study 1A: Distribution of Individual Participants' Prediction Strategies



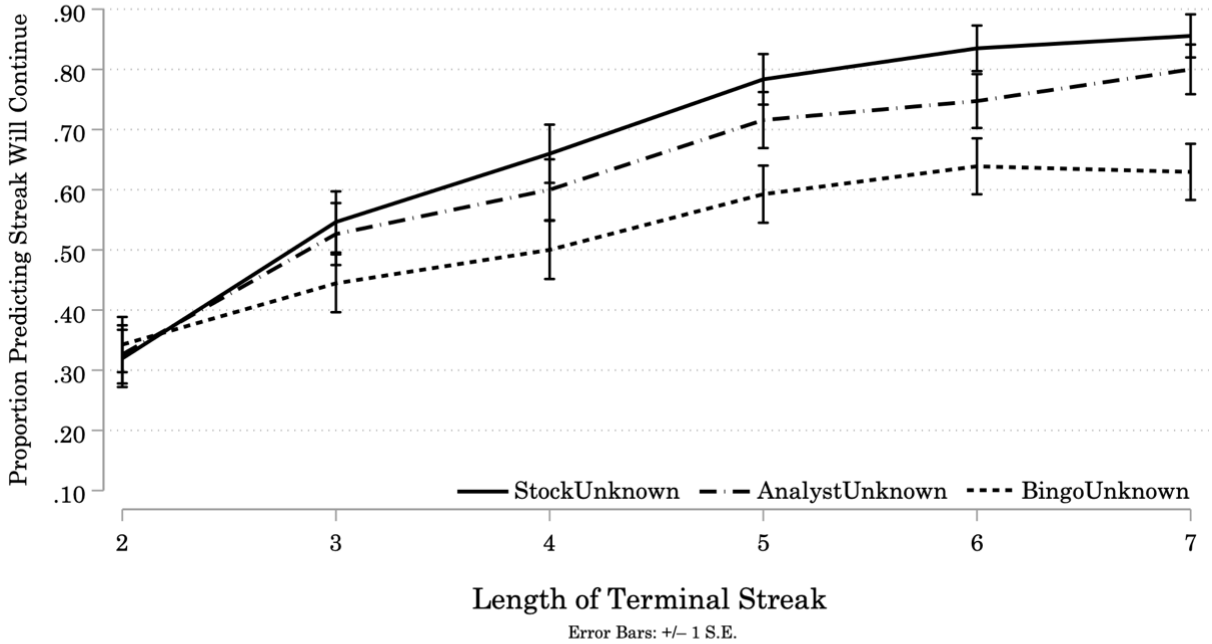
Note: Participants' individual prediction strategies in Study 1A. Histograms of frequencies for fitted slopes of participants' predictions over target stimuli ending in streaks of length 2–7 (gray bars), overlaid with normal density curves (dark gray lines). The distributions of slopes for participants in the AnalystUnknown, StockUnknown, and BingoUnknown Conditions are all centered above 5. With each additional signal added to the streak at the end of a target sequence, participants increased the probability they assigned to repetition of that streak by a little over 5%.

Results: Study 1B

Figure 3 presents the proportion of participants who predicted the 9th (next) outcome would repeat the streak at the end of each target sequence.

Figure 3

Study 1B: Proportion of Participants Predicting Repetition of Terminal Streaks, by Streak Length and Condition



Note: Results of Study 1B. Solid and patterned lines represent the proportion of participants in each Condition who predicted the next (9th) outcome will repeat the terminal streak at the end of each target sequence. Error bars represent +/– 1 standard error. Starting at streaks of length 4, more than 50% of participants in the StockUnknown and AnalystUnknown Conditions predicted that the next signal would repeat the streak. At streaks of length 5 and greater, more than 50% of participants in the BingoUnknown Condition predicted the next signal would repeat the streak. N = 300 (StockUnknown = 97; AnalystUnknown = 95; BingoUnknown = 108).

A one-way mixed ANOVA was conducted to test the effects of one between-subjects variable (Condition), and one within-subjects variable (Streak Length) on participants' predictions that the streak at the end of each target sequence would repeat. Participants'

predictions ("1" for repeat; "0" for reverse) were the focus of analysis.[21] Table 2 presents summary statistics by Condition and Streak Length.

Table 2

Study 1B: Proportion of Participants Predicting Repetition of Terminal Streaks, by Streak Length and Condition

		Proportion Who Predicted Streak Will Repeat					
Condition	N	Streak = 2	Streak = 3	Streak = 4	Streak = 5	Streak = 6	Streak = 7
AnalystUnknown	95	0.33 (0.47)	0.53 (0.50)	0.60 (0.49)	0.72 (0.45)	0.75 (0.43)	0.80 (0.40)
StockUnknown	97	0.32 (0.47)	0.55 (0.50)	0.66 (0.47)	0.78 (0.41)	0.84 (0.37)	0.86 (0.35)
BingoUnknown	108	0.34 (0.47)	0.44 (0.50)	0.50 (0.50)	0.59 (0.49)	0.64 (0.48)	0.63 (0.48)
All	300	0.33 (0.47)	0.51 (0.50)	0.59 (0.49)	0.70 (0.46)	0.74 (0.44)	0.76 (0.43)

Note: Standard deviations in parentheses.

There was a significant effect of Condition on participant predictions ($F(2, 297) = 5.62$, $p = 0.004$). Bonferroni-corrected pairwise comparisons revealed a significantly smaller proportion of participants in the BingoUnknown Condition predicted streaks would repeat than in the StockUnknown Condition (Mean Difference = –0.14, p = 0.004). However, there was not a significant difference between the proportion of participants predicting streaks would repeat in the AnalystUnknown Condition and in the BingoUnknown Condition (Mean Difference = 0.10, p = 0.092), or between participants in the AnalystUnknown Condition and in the StockUnknown Condition (Mean Difference= –0.05, p = 0.872). There was a significant main effect of Streak

[21] We present the results of the ANOVA here for ease of exposition, but it is not the appropriate analysis for these data. Because participants' responses are dichotomized (1, 0), the distributions of these responses are not normal. This violates the assumptions of the least-squares model used in ANOVA. Interested readers can find the results of a repeated measures binary logistic regression analysis of these data, as well as the binary choice data from Studies 2B and 3B, in the *Binary Logistic Regression Analyses* section of the Online Supplement. There are no differences between the substantive conclusions drawn from the ANOVA and the repeated measures binary logistic regression analyses for any of the studies presented in this article.

Length on predictions, ($F(4.52, 1343.13) = 57.74$, $p < 0.000$).[22] A higher proportion of participants predicted streaks would repeat as Streak Length increased. There was no significant interaction between Streak Length and Condition ($F(9.05, 1343.13) = 1.61$, $p = 0.107$). Participants responded to increases in Streak Length similarly across all three Conditions.

To get a sense of differences between individual participants' prediction strategies we took a look at the coefficients obtained from logistic regressions fitted to each participant's predictions over the target sequences ending in streaks of length 2–7.[23] For ease of exposition, we transformed the log odds coefficients obtained from these logistic regressions into the percent-change in the odds the participant predicts "repeat" for each unit increase in Streak Length.[24] A positive percent-change in the odds indicates that the likelihood the participant predicts repetition of the terminal streak increases as the length of that streak increases. A negative percent-change in the odds indicates the opposite strategy, that the likelihood the participant predicts repetition of the streak decreases as the length of that streak increases. Figure 4 presents the distribution of individual participants' strategies – the percent-change in the odds a participant predicts "repeat" for each unit increase in Streak Length.

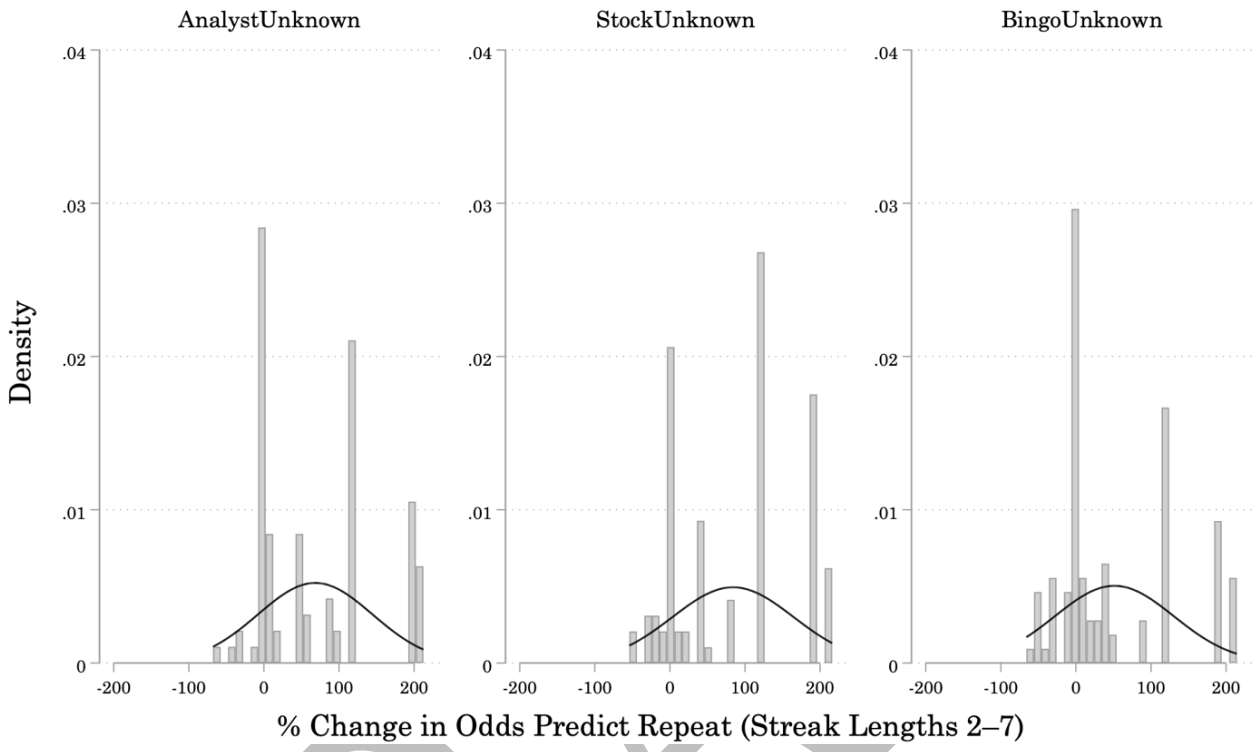
[22] Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(14) = 76.63$, $p < 0.000$); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.90$).

[23] Separation was observed while running participant-level logistic regressions on 127/300 participants' predictions over target sequences. Firth's procedure was applied to all of the participant-level logistic regressions to resolve the separation issue, producing less biased coefficients (for an explanation of this procedure, see Heinze & Schemper, 2002).

[24] We first exponentiate the coefficient to obtain the odds ratio, then we subtract 1 from the odds ratio and multiply by 100 to get the percent-change in the odds: $[(e^{\beta} - 1) \times 100]$. Example: The log odds that Participant A predicts "repeat" increase by 0.78 for each unit increase in Streak Length. The odds that this participant predicts a streak will repeat are $\exp(0.78) = 2.18$ times higher for each unit increase in Streak Length. The percent-change in the odds this participant predicts "repeat" is $[100 \times (2.18–1)] = 118\%$ for each unit increase in Streak Length.

Figure 4

Study 1B: Distribution of Individual Participants' Prediction Strategies



Note: Individual participants' prediction strategies in Study 1B. Logistic regressions were fitted to individual participants' predictions over target stimuli ending in streaks of length 2–7 to find the percent-change in the odds a participant predicts a streak will repeat for each unit increase in the length of that streak. Histograms of frequencies for different values of percent-changes (gray bars), overlaid with normal density curves (dark gray lines). The distributions of percent-changes for participants in the AnalystUnknown and StockUnknown Conditions are centered above 75%, and the distribution for participants in the BingoUnknown Condition is centered above 50%.

As in Study 1A, there was some heterogeneity in individual participants' prediction strategies. The distributions of percent-changes are centered above 75% in the AnalystUnknown and StockUnknown Conditions. The odds participants in these Conditions predict repetition for

a streak of length x are about 75% higher than the odds a participant in these Conditions predict repetition for a streak of length $x - 1$. The distribution of values is centered above 50% in the BingoUnknown Condition. A minority of participants exhibit negative percent-changes in the odds they will predict "repeat" as Streak Length increases: 8% in the AnalystUnknown Condition, 10% in the StockUnknown Condition, and 16% in the BingoUnknown Condition.

Discussion: Studies 1A and 1B

When given no information about the distribution of base rates, the majority of participants in Study 1A and Study 1B increased their expectations that a streak would repeat as the length of that streak increased (hot hand). A small number of participants in each study decreased their expectations that a streak would repeat as the length of that streak increased (gambler's fallacy). We interpret these findings as providing support for Rabin's and Burns's suggestions that a reasonable process of updating beliefs about a generator's base rate leads an observer's expectation of repetition to increase with Streak Length. The ordering of participants' predictions does not fit the hypothesis that participants facing an intentional actor are more prone to believe in a momentum or hot hand process than participants facing a random mechanical device or a market. The only statistically reliable difference was between predictions made by participants facing the Stock generator and those facing the Bingo generator. Differences in the slopes of individual participants' prediction curves suggest that a reversal bias, perhaps attributable to gambler's fallacy thinking, is likeliest to occur when the generator is described as a random mechanism.

In terms of differences across response types, it does not seem that participants were more likely to exhibit a positive recency bias (predicting repetition) in Study 1B than in Study 1A. In fact, the opposite seems to be the case. In order to compare the responses of participants

in Studies 1A and 1B, participants' responses in Study 1A were dichotomized. All predictions equal to 50% were dropped (12 predictions across 8 participants). Predictions higher than 50% were coded as "1" (streak will repeat), and predictions lower than 50% were coded as "0" (streak will reverse). Our method for recoding the results from Study 1A is not arbitrary, the 50% threshold seems like a neutral and meaningful choice. But, it may not accurately reflect participants' decision rule, and so the following discussion is speculative.

We performed Welch's t-test for unequal variances to compare the proportion of participants predicting streaks would repeat in Study 1A to the proportion of participants predicting streaks would repeat in Study 1B.[25] A significantly smaller proportion of participants in each Condition of Study 1B predicted streaks would repeat than in each corresponding Condition of Study 1A ($\text{Bingo}_{\text{STUDY1B}} - \text{Bingo}_{\text{STUDY1A}} = -0.14$, *s.e.* = 0.03, $p < 0.000$; $\text{Analyst}_{\text{STUDY1B}} - \text{Analyst}_{\text{STUDY1A}} = -0.11$, *s.e.* = 0.03, $p < 0.000$; $\text{Stock}_{\text{STUDY1B}} - \text{Stock}_{\text{STUDY1A}} = -0.13$, *s.e.* = 0.03, $p < 0.000$). A greater proportion of participants in the BingoUnknown and StockUnknown Conditions predicted streaks would repeat in Study 1A (Continuous Response) than in Study 1B (Binary Response), while the proportion of participants predicting repetition was about the same in the AnalystUnknown Condition of both Study 1A and Study 1B.

Section IV: Studies 2A and 2B

In Studies 2A and 2B, we provided explicit instructions that each generator produces each type of outcome at a stationary rate of .50. Our goal was to determine whether provision of a stationary base rate causes participants to exhibit the U-shaped judgment pattern predicted by Rabin (2002), initially *decreasing* expectations of streak repetition, and then *increasing* expectations of repetition as Streak Length increases. We were also interested in whether

[25] We selected Welch's *t*-test because it does not assume equal variances, and is robust to large differences in sample sizes across groups. Note that there were 300 participants in Study 1B, but only 144 participants in Study 1A.

participants respond to stationary base rate information differently depending on the qualitative description of the generator (e.g. does knowledge of a stationary .50 base rate produce different judgment patterns when the generator is described as a bingo cage versus an investment analyst?).

Participants

156 participants (M_{AGE} = 35.57, SD_{AGE} = 10.25, N_{FEMALE} = 69) were recruited for Study 2A. Participants took 18.15 minutes on average (SD = 9.67) to complete the Study 2A. 301 participants (M_{AGE} = 35.58, SD_{AGE} = 12.51, N_{FEMALE} = 141) were recruited for Study 2B. Participants took 18.57 minutes on average (SD = 9.47) to complete Study 2B. Participants in both studies were paid $2.50 upon approval of their completed tasks.

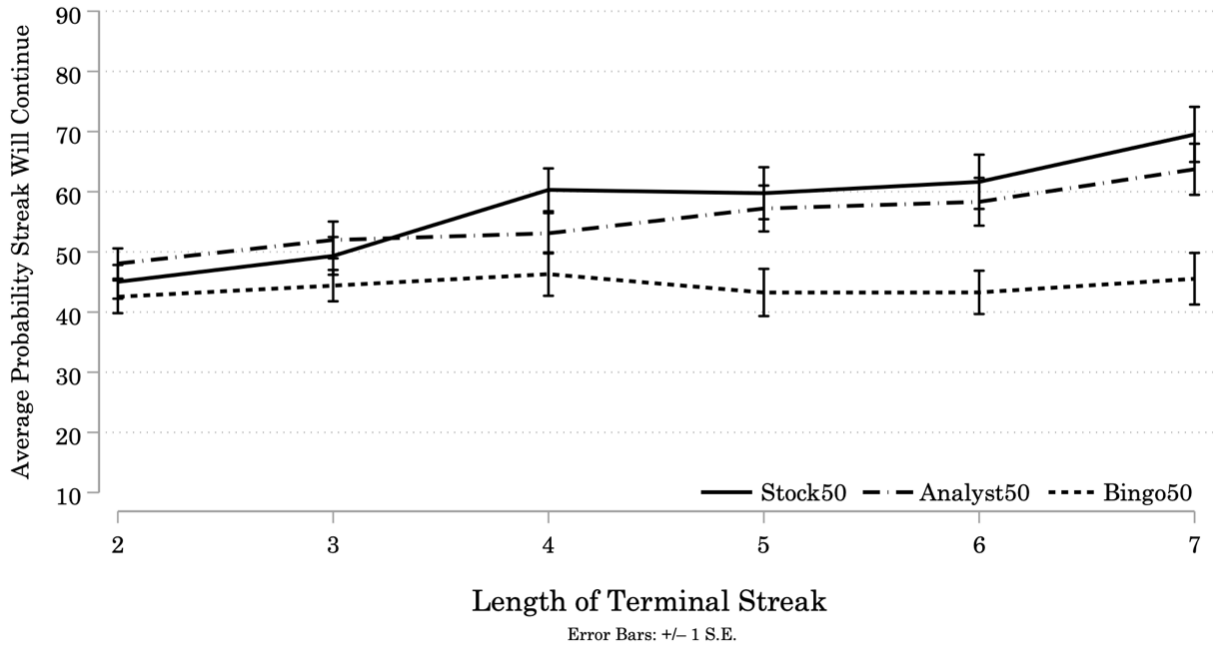
Method

Participants in Study 2A made their predictions using a continuous numerical probability scale, and participants in Study 2B made their predictions as a dichotomous choice. In each study, participants were randomly assigned to one of three experimental conditions, defined by the description and base rate of the generator. In the Bingo50 Condition, the events in each sequence were described as draws (Red/Blue) made *with replacement* by a mechanical bingo machine from a cage containing exactly 50 red and 50 blue balls. In the Analyst50 Condition, the events in each sequence were described as quarterly changes (Up/Down) in the value of a particular investment analyst's portfolio. Participants were instructed that all analysts they encounter in the experiment are of *average* skill level, and that their portfolios increase in value 50% of the time. In the Stock50 Condition, the events in each sequence were described as quarterly changes (Up/Down) in a particular company's stock price. Participants were instructed that all companies they encounter in the experiment have an *average* performance level, and that

their stock prices increase in value 50% of the time. On each trial, participants were instructed that they were viewing a *new* sequence to reinforce the idea that they were not seeing a continuation of the previous trial's sequence. We also reminded participants of the stationary .50 rate of each generator on each trial. Aside from these differences in the instructions, the experimental procedure was identical to that used Studies 1A and 1B.

Figure 5

Study 2A: Average Probability Participants Assigned to Repetition of Terminal Streaks, by Streak Length and Condition



Note: Results of Study 2A. Solid and patterned lines represent the average probability participants assigned to repetition of the terminal streak at the end of each target sequence. Error bars represent +/– 1 standard error. Starting at streaks of length 4, participants in the Analyst50 and Stock50 Conditions assigned greater than 50% probability to repetition of the streak, and their expectations of repetition increased with Streak Length. Participants in the Bingo50 Condition consistently assigned less than 50% probability to repetition of the streak. Participants

in the Bingo50 Condition did not consistently increase or decrease their expectations of repetition with Streak Length. N = 156 (Stock50= 48; Analyst50 = 52; Bingo50 = 56).

Results: Study 2A

Figure 5 presents the average probability participants assigned to repetition of the terminal streak at the end of each target sequence. Predictions made by participants faced with an intentional actor (Analyst50 Condition) were *not* significantly different than predictions made by participants faced with a market process (Stock50 Condition). For streaks of length 2 or 3, participants made similar predictions across all three Conditions. Starting at streaks of length 4, participants in the Analyst50 and Stock50 Conditions both assigned greater than 50% probability to the next signal repeating the streak, and increased the probability they assigned to repetition as Streak Length increased. But, participants in the Bingo50 Condition continued to assign less than 50% probability to the next (9th) outcome repeating the terminal streak, and neither increased nor decreased the probability they assigned to repetition as Streak Length increased.

Table 3 presents the summary statistics for Study 2A, by Condition and Streak Length. We conducted a one-way mixed ANOVA to test the effects of one between-subjects variable (Condition), and one within-subjects variable (Streak Length) on participants' predictions that the streak at the end of each target sequence would *repeat*. There was a significant effect of Condition on participant predictions ($F(2, 153) = 7.46$, $p = 0.001$). Bonferroni-corrected pairwise comparisons revealed that the mean predictions made by participants in the Bingo50 Condition were significantly lower than the mean predictions made by participants in the Analyst50 Condition (Mean Difference = −11.19, $p = 0.009$), and by participants in the Stock50 Condition (Mean Difference = −13.38, $p = 0.002$). There was no significant difference between the predictions made by participants in the Analyst50 and Stock50 Conditions (Mean Difference

= −2.20, p = 1.000). Participants faced with a random mechanism (balls drawn from a bingo cage) were more likely to predict reversals across all target sequences, and participants faced with an intentional actor (investment analyst) or a market process (publicly-traded company) were more likely to predict repetition of terminal streaks of length 4 or greater.[26]

Table 3

Study 2A: Average Probability Participants Assigned to Repetition of Terminal Streaks, by Streak Length and Condition

		Average Probability (Streak Will Repeat)					
Condition	N	Streak = 2	Streak = 3	Streak = 4	Streak = 5	Streak = 6	Streak = 7
Analyst50	52	48.06 (18.15)	51.98 (22.04)	53.10 (24.15)	57.21 (27.51)	58.33 (28.65)	63.73 (30.62)
Stock50	48	45.02 (19.43)	49.33 (21.75)	60.31 (24.67)	59.75 (29.93)	61.65 (31.25)	69.52 (31.75)
Bingo50	56	42.54 (20.36)	44.39 (19.51)	46.30 (26.97)	43.25 (29.32)	43.27 (26.91)	45.54 (32.09)
All	156	45.14 (19.37)	48.44 (21.18)	52.88 (25.83)	52.98 (29.67)	53.94 (29.82)	58.98 (32.97)

Note: Standard deviations in parentheses.

There was a significant main effect of Streak Length on predictions, (F(3.91, 598.01) = 10.07, p < 0.000).[27] Longer streak lengths were assigned higher probability of continuing. There was a significant interaction between Streak Length and Condition (F(7.82, 598.01) = 2.56, p = 0.010). Bonferroni-corrected pairwise comparisons revealed no significant differences between predictions made by participants in each of the three Conditions for streaks of length 2 or 3. For streaks of length 4, there was no significant difference between predictions made by participants in the Analyst50 and Bingo50 Conditions (Mean Difference = 6.79, p = 0.499), but Stock50

[26] A one-sample, one-tailed t-test also confirmed that the mean of participants' predictions across all Streak Lengths in the Bingo50 Condition was significantly less than 50% (t = –4.06, p < 0.000).

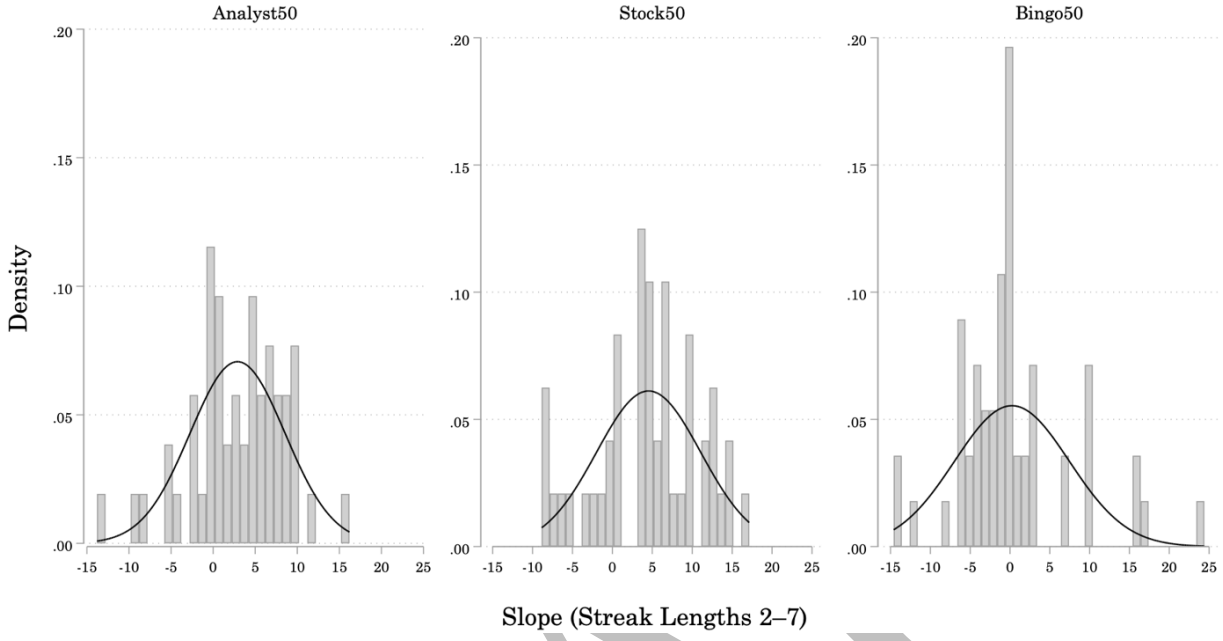
[27] Mauchly's test indicated that the assumption of sphericity had been violated (χ^2(14) = 89.94, p < 0.000); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = 0.82).

participants' predictions were significantly higher than those of Bingo50 participants (Mean Difference = 14.01, $p = 0.017$). For streaks of length 5, 6, and 7, participants in the Analyst50 and Stock50 Conditions assigned significantly higher probability to repetition of the terminal streak than did participants in the Bingo50 Condition. Participants in the Analyst50 and Stock50 Conditions seemed to update their estimates of the generator's base rate as Streak Length increased. This updating behavior is similar to, but more conservative than, what we observed in Study 1A. Participants in the Bingo50 Condition seemed to follow a different judgment process. Their predictions are consistently lower than 50% (but did not decrease with Streak Length).

To understand participants' individual prediction strategies, we again take a look at the slopes of linear regressions fitted to each participants' predictions over the 6 target sequences (Figure 6). Individual participants employed similar, albeit more conservative, updating strategies in the Analyst50 and Stock50 Conditions (Study 2A) as they did in the AnalystUnknown and StockUnknown Conditions (Study 1A). But, we do see a difference between the prediction strategies employed in the Bingo50 Condition and the BingoUnknown Condition (compare Figures 6 and 2). There were more extreme outliers in both the left- and right-tails of the distribution in the Bingo50 Condition in Study 2A than there were in the BingoUnknown Condition of Study 1A. The predictions made by 54% of participants in the Bingo50 Condition in Study 2A exhibited a negative slope, compared to only 20% in the BingoUnknown Condition of Study 1A.

Figure 6

Study 1B: Distribution of Individual Participants' Prediction Strategies



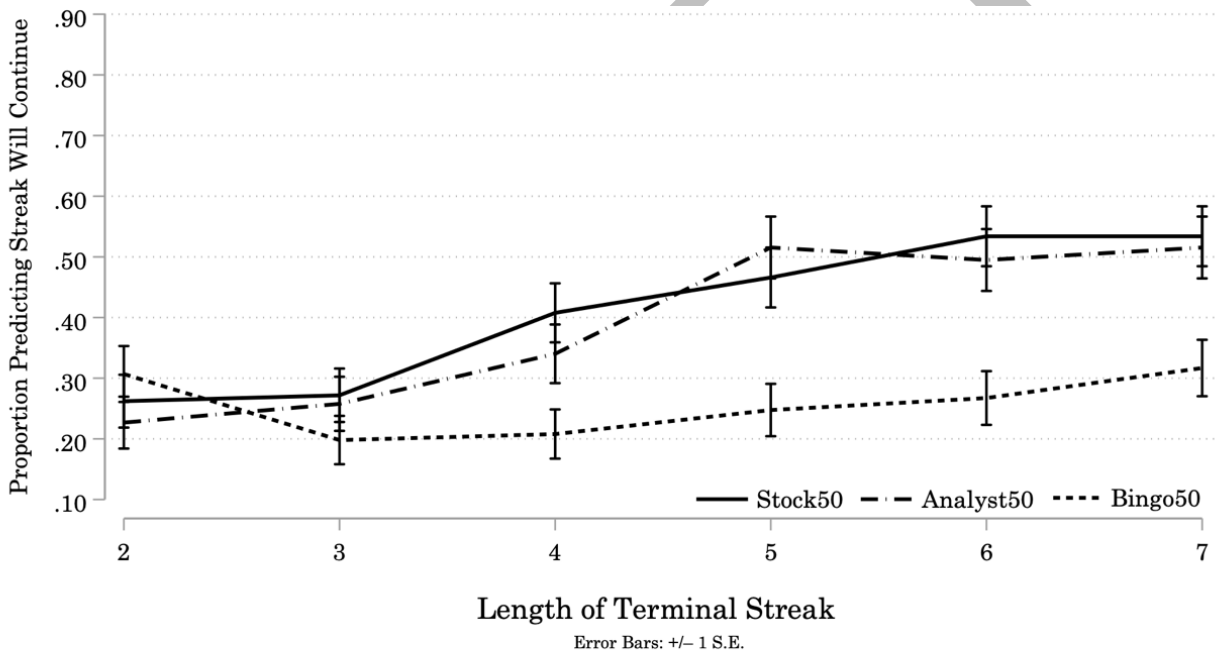
Note: Individual participants' prediction strategies in Study 2A. Histograms of frequencies for fitted slopes of participants' predictions over target stimuli ending in streaks of length 2–7 (gray bars), overlaid with normal density curves (dark gray lines). The distributions of slopes for participants in the Analyst50 and Stock50 Conditions are centered between 2.5 and 5. With each additional signal added to the streak at the end of a target sequence, participants increased the probability they assigned to repetition of that streak by 2.5 to 5%. The distribution of slopes for participants in the Bingo50 Condition is centered above zero.

If we focus on the average results, we would conclude that participants given a stationary base rate for a random generator (Bingo50 Condition) show a bias to predict reversals, consistent with gambler's fallacy reasoning. And, the expectation of reversal appears constant across Streak Lengths. In spite of the explicit instructions about a stationary base rate, participants faced with a market process (Stock50 Condition) or an intentional agent (Analyst50 Condition)

generator seem to update their estimates of the base rate as the length of the streak increases to the point that the base rate provided in the instructions becomes implausible. (We interpret the fact that participants updated more conservatively in Study 2A than in Study 1A as an indication that the specific base rate instructions they received in 2A were at least somewhat compelling).

Figure 7

Study 2B: Proportion of Participants Predicting Repetition of Terminal Streaks, by Streak Length and Condition



Note: Results of Study 2B. Solid and patterned lines represent the proportion of participants in each Condition who predicted repetition of the terminal streak at the end of each target sequence. Error bars represent +/– 1 standard error. Across streaks of length 2–4, fewer than 50% of participants in all three Conditions predicted that the streak would repeat. The proportion of participants predicting repetition in the Analyst50 and Stock50 Conditions increased across Streak Length 2–5, before leveling off at around 50%. In contrast, only 20% to 30% of

participants in the Bingo50 Condition predicted streaks of any length would repeat. N = 301 (Stock50 = 103; Analyst50 = 97; Bingo50 = 101).

Results: Study 2B

Figure 7 shows the proportion of participants who predicted the streak at the end of each target sequence would repeat. Across streaks of length 2–4, fewer than 50% of participants in all three Conditions predicted the streak would repeat. The proportion of participants predicting repetition in the Analyst50 and Stock50 Conditions increased across Streak Lengths 2–5, before leveling off at around 50%. In contrast, only 20% to 30% of participants in the Bingo50 Condition predicted repetition across all Streak Lengths.

We conducted a one-way mixed ANOVA to test the effects of one between-subjects variable (Condition), and one within-subjects variable (Streak Length) on participants' predictions that the streak at the end of each target sequence would *repeat*. Participants' predictions ("1" for *repeat*; "0" for *reverse*) were the focus of analysis.[28] Table 4 presents summary statistics by Condition and Streak Length. There was a significant effect of Condition on participant predictions ($F(2, 298) = 7.09$, $p = 0.001$). Bonferroni-corrected pairwise comparisons revealed a significantly smaller proportion of participants in the Bingo50 Condition predicted streaks would repeat than in the Stock50 Condition (Mean Difference = –0.16, $p = 0.002$) and in the Analyst50 Condition (Mean Difference = –0.13, $p = 0.010$). However, there was *not* a significant difference between the proportion of participants predicting streaks would repeat in the Analyst50 and Stock50 Conditions (Mean Difference = –0.02, $p = 1.000$).

[28] The results of a repeated measures binary logistic regression analysis of these results can be found in the section of the Online Supplement titled *Binary Logistic Regression Analyses*.

There was a significant main effect of Streak Length on predictions, ($F(4.47, 1332.46) = 17.86$, $p < 0.000$).[29] A higher proportion of participants predicted streaks would repeat as Streak Length increased. The interaction between Streak Length and Condition was also significant ($F(8.94, 1332.46) = 3.84$, $p < 0.000$). In the Stock50 and Analyst50 Conditions, the proportion of participants predicting a streak would repeat increased over Streak Lengths 2 through 5. In the Bingo50 Condition, the proportion of participants predicting a streak would repeat did not increase with Streak Length.

Table 4

Study 2B: Proportion of Participants Predicting Repetition of Terminal Streaks, by Streak Length and Condition

		Proportion Who Predicted Streak Will Repeat					
Condition	N	Streak = 2	Streak = 3	Streak = 4	Streak = 5	Streak = 6	Streak = 7
Analyst50	97	0.23 (0.42)	0.26 (0.44)	0.34 (0.47)	0.52 (0.50)	0.50 (0.50)	0.52 (0.50)
Stock50	103	0.26 (0.44)	0.27 (0.44)	0.41 (0.49)	0.47 (0.50)	0.53 (0.50)	0.53 (0.50)
Bingo50	101	0.31 (0.46)	0.20 (0.40)	0.21 (0.41)	0.25 (0.43)	0.27 (0.44)	0.32 (0.47)
All	301	0.27 (0.44)	0.24 (0.43)	0.32 (0.47)	0.41 (0.50)	0.43 (0.50)	0.46 (0.50)

Note: Standard deviations in parentheses.

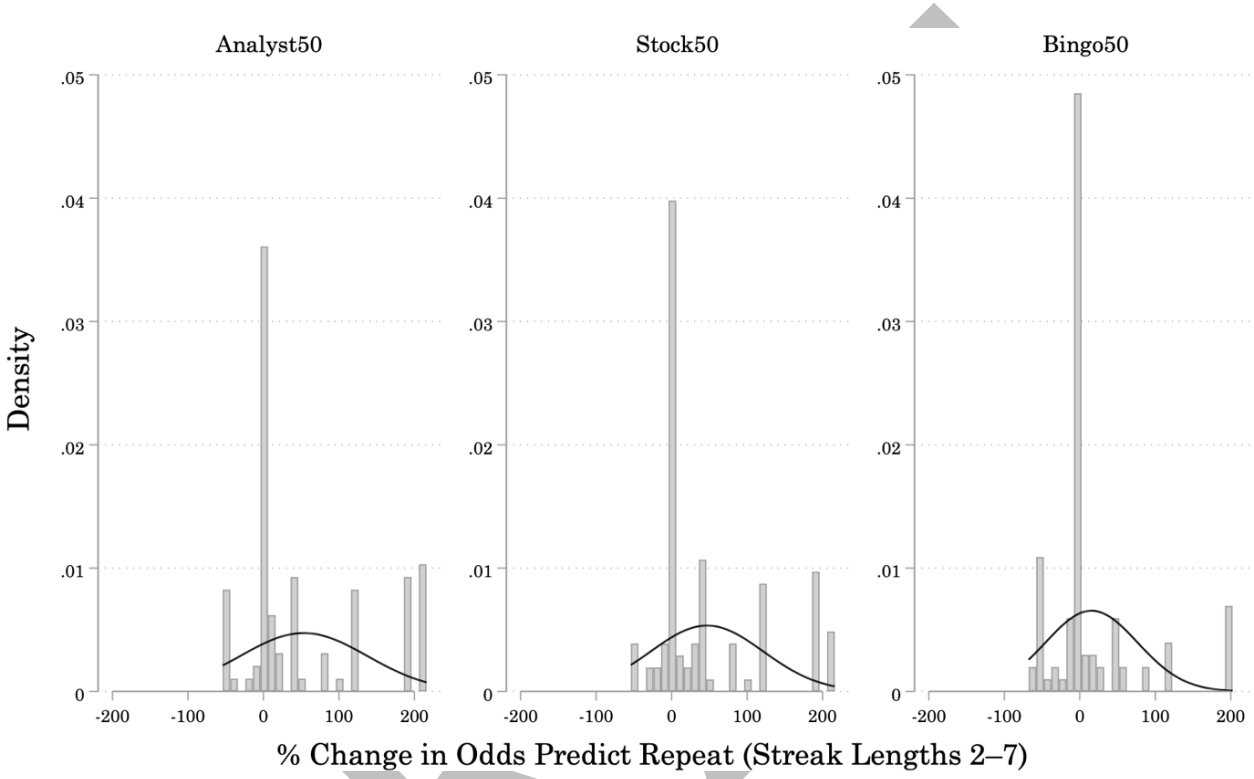
To understand participants' individual prediction strategies, we again fitted logistic regressions to each participants' predictions for the six target sequences ending in a streak.[30] We then transformed each participant's coefficient from the log odds to the percent-change in the

[29] Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(14) = 79.88$, $p < 0.000$); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.89$).
[30] Separation was observed while running participant-level logistic regressions on 83/301 participants' predictions over target sequences. Firth's procedure was applied to all of the participant-level logistic regressions to resolve the separation issue.

odds of predicting repetition for each unit increase in Streak Length. Figure 8 presents the distributions of these values in each Condition.

Figure 8

Study 2B: Distribution of Individual Participants' Prediction Strategies



Note: Individual participants' prediction strategies in Study 2B. Logistic regressions were fitted to participants' predictions over target stimuli ending in streaks of length 2–7 to find the percent-change in the odds a participant predicts a streak will repeat for each unit increase in the length of that streak. Histograms of frequencies for different values of percent-changes (gray bars), overlaid with normal density curves (dark gray lines). The distributions of percent-changes for participants in the Analyst50 and Stock50 Conditions are centered above 50%, and the distribution for participants in the Bingo50 Condition is centered above 15%.

The distributions of percent-change values are centered above 50% in the Analyst50 and Stock50 Conditions, and the distribution of values is centered above 15% in the Bingo50

Condition. The odds that a participant in the Bingo50 Condition would predict repetition for a streak of length x were about 15% higher than the odds a participant in this Condition would predict repetition for a streak of length $x - 1$. A minority of participants exhibit negative percent-changes in the odds they predict "repeat" as Streak Length increases: 14% in the Analyst50 Condition, 13% in the Stock50 Condition, and 26% in the Bingo50 Condition.

Discussion: Studies 2A and 2B

The results from Studies 2A and 2B were surprising to us. The results were not consistent with a gambler's fallacy pattern of predictions, as expectations of repetition did not consistently decrease across Streak Lengths. Neither were they perfectly consistent with Rabin's (2002) model that posits an initial *decrease*, followed by an *increase* in expectations of repetition as Streak Length increases. Participants facing all three generators showed a slight bias toward reversal for streaks of length 2. For streaks of length 4 through 7, participants faced with an intentional actor (Analyst50 Condition) or a market process (Stock50 Condition) slightly increased their expectations of repetition, while participants faced with a random mechanical generator (Bingo50 Condition) continued to show a bias toward reversal across all Streak Lengths (however, expectations of reversal did not *increase* with Streak Length). Our inclination is to explain these results by proposing different mental models of the random device generator versus the intentional actor and market generators. Participants might have interpreted the .50 performance rate of the companies and investment analysts as the *average* of two states (e.g. striving versus slacking). This would make transitions between states (high versus low performance) more *a priori* plausible for these generators than for the bingo cage, which could not transition between states because it always contained exactly 50 red and 50 blue balls.

We checked for differences across response types by dichotomizing the predictions made by participants in Study 2A. Welch's *t*-test for unequal variances revealed that a significantly smaller proportion of participants in each Condition of Study 2B predicted streaks would repeat than in each Condition of Study 2A ($\text{Bingo}_{\text{STUDY2B}} - \text{Bingo}_{\text{STUDY2A}} = -0.11$, *s.e.* = 0.03, $p < 0.000$; $\text{Analyst}_{\text{STUDY2B}} - \text{Analyst}_{\text{STUDY1A}} = -0.23$, *s.e.* = 0.03, $p < 0.000$; $\text{Stock}_{\text{STUDY2B}} - \text{Stock}_{\text{STUDY1A}} = -0.17$, *s.e.* = 0.04, $p < 0.000$). Participants were not more likely to exhibit a preference for streaks to repeat when asked to provide a Binary Response (Study 2B) than when asked to provide a Continuous Response (Study 2A). Once again, the opposite seems to be the case. A greater proportion of participants in the Analyst50 and Stock50 Conditions predicted streaks would repeat in Study 2A than in Study 2B. The proportion of participants predicting repetition in the Bingo50 Condition was similar in both studies.

Section V: Studies 3A and 3B

In Studies 3A and 3B, we specified a precise distribution of possible rates for each generator, giving us more control over the level of uncertainty in participants' prior beliefs about each generator's base rate. Participants were told that there were exactly three possible rates for each generator (.25, .50 or .75), and that each of these rates was equally likely to generate the sequence on each trial (the probability that a given sequence was generated by each of the three rates is 1/3). We speculated that explicitly specifying a precise distribution of rates (three alternative performance states) for the company, investment analyst, and bingo cage may increase consistency across participants' mental models of these generators. We hypothesized that such consistency may lead the pattern of predictions made by participants facing each of these generators to converge. Specifying the distribution of possible rates also allows us to use a Bayesian updating model to estimate the posterior probability that the terminal streak at the end

of each target sequence will repeat. We can then use these posterior estimates as a "rational" benchmark, against which we'll compare participants' predictions.

Participants

150 participants (M_{AGE} = 34.09, SD_{AGE} = 10.06, N_{FEMALE} = 74) were recruited for Study 3A. Participants took 18.76 minutes on average (SD = 9.45) to complete Study 3A. They were paid $2.50 upon approval of their completed tasks. 300 participants (M_{AGE} = 36.83, SD_{AGE} = 11.86, N_{FEMALE} = 159) were recruited for Study 3B. Participants took 19.44 minutes on average (SD = 9.49) to complete Study 3B. Participants in both studies were paid $2.50 upon approval of their completed tasks.

Method

Participants in Study 3A indicated their predictions using a continuous sliding scale, and participants in Study 3B indicated their predictions by making a dichotomous choice. In each study, participants were randomly assigned to one of three experimental conditions, defined by the description and base rate of the generator. In the Bingo25-50-75 Condition, the events in each sequence were described as draws (Red/Blue) made *with replacement* by a mechanical bingo machine from one of three cages, each having a different ratio of red to blue balls (25:75, 50:50, and 75:25). On each trial, the machine randomly selects one of the cages with equal probability, and then draws 8 outcomes from the selected cage. In the Analyst25-50-75 Condition, the events in each sequence were described as quarterly changes (Up/Down) in the value of a particular investment analyst's portfolio. Analysts were equally likely to have each of three skill levels, indicating the proportion of the time that their portfolios increase in value: Bad (.25), Average (.50), and Good (.75). In the Stock25-50-75 Condition, the events in each sequence were described as quarterly changes (Up/Down) in a particular company's stock price.

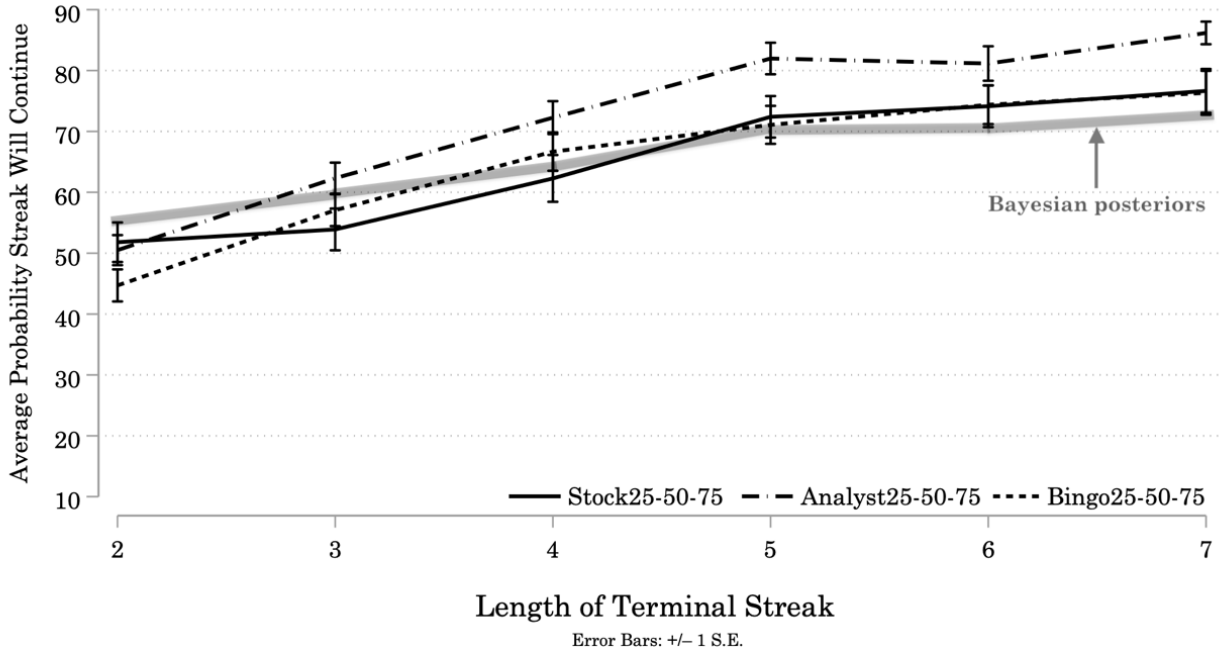
Companies were equally likely to have each of three performance levels, indicating the proportion of the time that their stock price increased: Bad (.25), Average (.50), and Good (.75). On each trial, participants were instructed that they were viewing a *new* sequence to reinforce the idea that they were not seeing a continuation of the previous trial's sequence. We also reminded participants of the distribution of possible rates, and that these rates were equally likely to have produced the present sequence on each trial. Aside from these differences in the instructions, the experimental procedure was identical to that used Studies 1A, 1B, 2A, and 2B.

Results: Study 3A

Figure 9 presents the average probabilities participants assigned to repetition of the streak at the end of each target sequence. Participants in all three Conditions assigned greater than 50% probability to repetition of streaks of length 3 or greater. Predictions made by participants in all three Conditions follow a reasonable updating pattern as Streak Length increased, up to the point where the highest possible rate (.75) becomes the most likely (when Streak Length = 7). Study 3A was designed to allow us to compare posteriors based on a rational Bayesian updating process with participants' posterior beliefs. In Figure 9, these Bayesian posteriors are superimposed (transparent gray line) over the predictions made by participants in each Condition. Participants in the Analyst25-50-75 Condition appear to overreact to streaks of length 4 or longer, assigning probabilities that are about 10 points higher than the Bayesian posteriors. Although the Analyst25-50-75 predictions are not significantly higher than those in the other two Conditions, this is the one feature in our results that could be interpreted as hinting that intentionality increases expectations of repetition above and beyond the results of a rational belief-updating process.

Figure 9

Study 3A: Average Probability Participants Assigned to Repetition of Terminal Streaks, by Streak Length and Condition



Note: Results of Study 3A. Solid and patterned lines represent the average probability participants in each Condition assigned to repetition of the streak at the end of each target sequence. Error bars represent +/– 1 standard error. The transparent gray bar represents the average Bayesian posterior probabilities of terminal streaks repeating, conditional on the pattern of outcomes in each target sequence. Participants' expectations of repetition increased with Streak Length. Probabilities assigned by participants in the Bingo25-50-75 and Stock25-50-75 Conditions converged to the maximum possible rate (.75) by Streak Length 6. Probabilities assigned by participants in the Analyst25-50-75 Condition exceeded the maximum possible rate starting with Streaks Length 5, and increased beyond the maximum rate across Streak Lengths 6 and 7. N = 150 (Stock25-50-75 = 50; Analyst25-50-75 = 50; Bingo25-50-75 = 50).

Table 5 presents summary statistics for Study 3A, by Condition and Streak Length. We conducted a one-way mixed ANOVA to test the effects of one between-subjects variable (Condition), and one within-subjects variable (Streak Length) on participants' predictions that the streak at the end of each target sequence would *repeat*. There was a significant effect of Condition on participants' predictions ($F(2, 147) = 3.59$, $p = 0.030$). However, Bonferroni-corrected pairwise comparisons reveal that differences between Conditions are only marginally significant. Participants in the Analyst25-50-75 Condition assigned slightly higher probabilities than those in the Bingo25-50-75 (Mean Difference = 7.36, $p = 0.061$) and Stock25-50-75 (Mean Difference = 7.20, $p = 0.069$) Conditions.

Table 5

Study 3A: Average Probability Participants Assigned to Repetition of Terminal Streaks, by Streak Length and Condition.

		Average Probability (Streak Will Repeat)					
Condition	N	Streak = 2	Streak = 3	Streak = 4	Streak = 5	Streak = 6	Streak = 7
Analyst25-50-75	50	50.50 (17.46)	62.32 (18.03)	72.26 (19.13)	81.98 (18.30)	81.16 (20.04)	86.18 (13.12)
Stock25-50-75	50	51.80 (22.95)	53.90 (24.25)	62.28 (27.15)	72.40 (24.11)	74.14 (24.32)	76.66 (25.44)
Bingo25-50-75	50	44.70 (18.65)	57.06 (18.49)	66.68 (22.27)	71.08 (22.06)	74.38 (24.32)	76.34 (25.57)
All	150	49.00 (19.94)	57.76 (20.61)	67.07 (23.30)	75.15 (22.02)	76.56 (22.48)	79.73 (22.48)

Note: Standard deviations in parentheses.

Recall that in Study 1A, Participants assigned significantly higher probability to repetition of streaks in the StockUnknown than to streaks in the BingoUnknown Condition. In Study 3A, there was no significant difference between predications made by participants in the Bingo25-50-75 and Stock25-50-75 Conditions (Mean Difference = –0.16, $p = 1.000$). We conjecture that the pattern we observed in Study 1A was driven by differences in participants'

mental models of the market and bingo cage generators. When identical distributions were specified for these generators in Study 3A, participants based their predictions about these generators on the same set of prior beliefs, resulting in identical prediction patterns.

There was a significant main effect of Streak Length on predictions, ($F(4.41, 647.80) = 82.32$, $p < 0.000$).[31] Longer Streak Lengths were assigned higher probability of repeating by participants in all three Conditions. The interaction between Streak Length and Condition was not significant ($F(8.81, 647.80) = 1.28$, $p = 0.247$). We interpret these results as evidence that an increasing expectation of repetition (hot hand) arises from uncertainty over the base rate of the generator. Participants' predictions converge toward the rate with the highest posterior probability, conditional on the sequence of signals they observed.

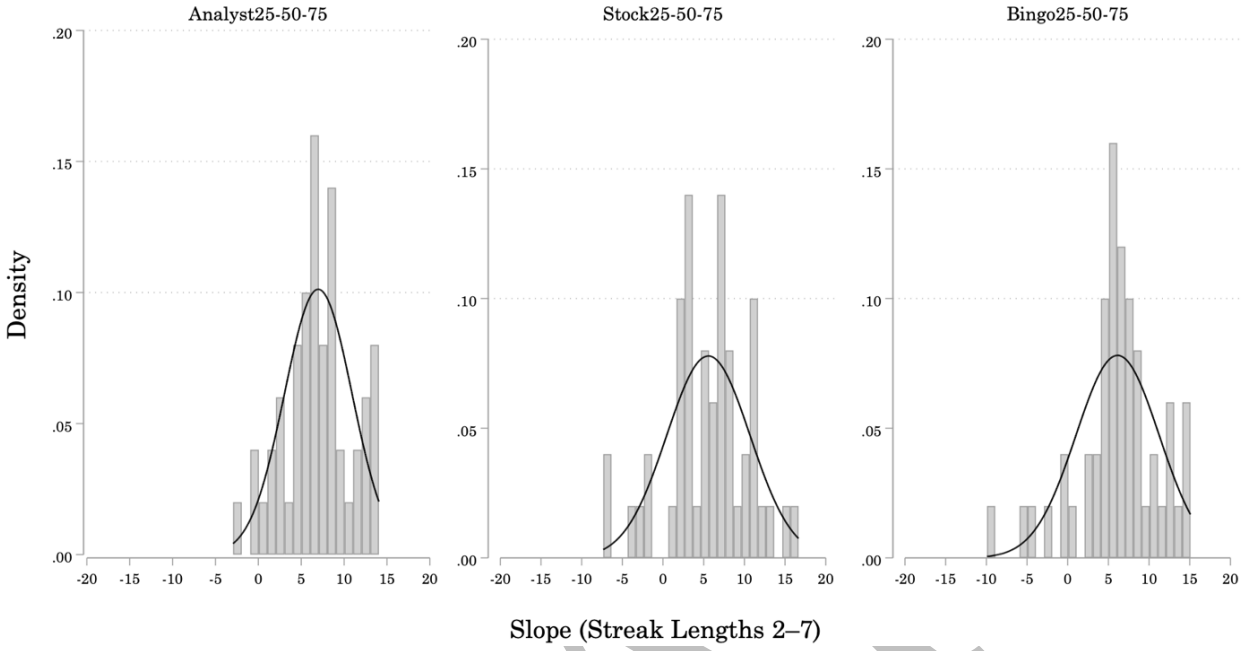
Figure 10 summarizes participants' prediction strategies. The distribution of slopes from regressions fitted to each participant's predictions over the target sequences are centered just above 5 in all three Conditions. For each unit increase in Streak Length, participants increased the probability they assigned to repetition of that streak by a little over 5%. There is more heterogeneity in prediction strategies used by participants in the Stock25-50-75 Condition than by those in the Analyst25-50-75 and Bingo25-50-75 Conditions.[32] There are also more extreme negative "outlier" strategies – participants whose expectations of repetition *decrease* as Streak Length increases – in the Bingo25-50-75 Condition than in the other two Conditions.

[31] Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(14) = 52.36$, $p < 0.000$); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.88$).

[32] We speculate this pattern results from the prevalence of contradictory beliefs about the stock market.

Figure 10

Study 3A: Distribution of Individual Participants' Prediction Strategies



Note: Individual participants' prediction strategies in Study 3A. Histograms of frequencies for slopes of linear regressions fitted to each participant's predictions over target stimuli (gray bars), overlaid with normal density curves (dark gray lines). Distributions of individual prediction slopes are centered just above 5 in all three Conditions. For each unit increase in the length of the terminal streak, participants increased the probability they assigned to the event that streak would repeat by a little over 5%.

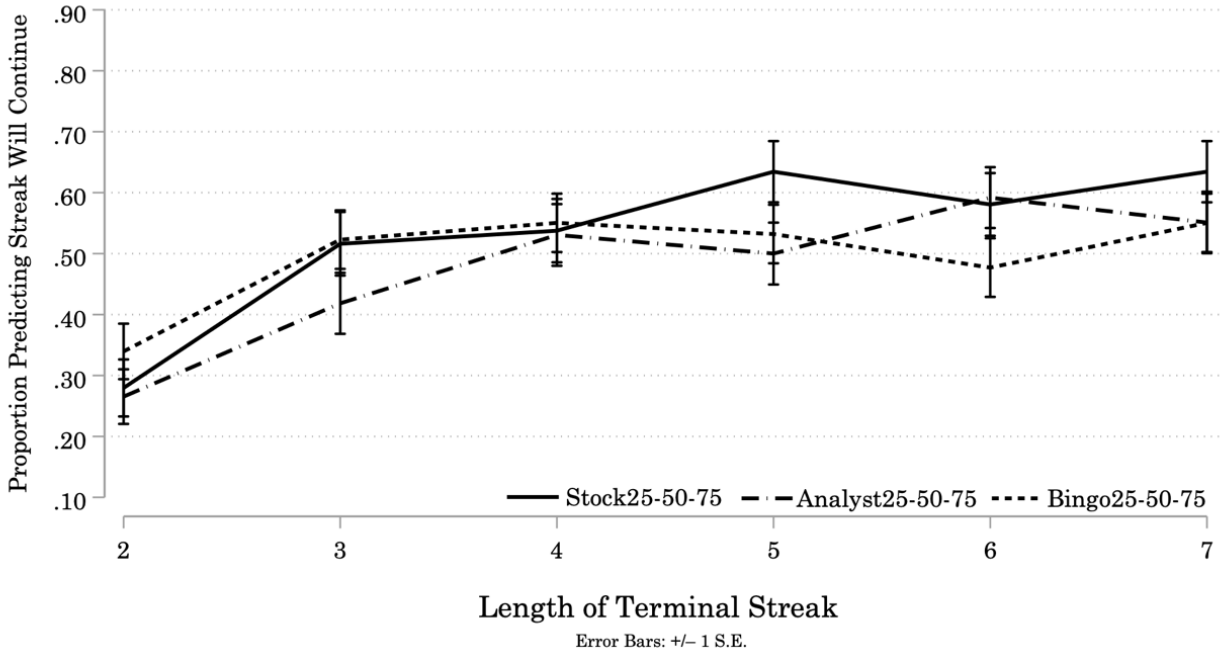
Results: Study 3B

Figure 11 shows the proportion of participants who predicted the streak at the end of each target sequence would *repeat*. In all three Conditions, the proportion of participants predicting repetition of the terminal streak increased between Streak Lengths 2 and 3. Across Streak Lengths 4 through 7, the proportion of participants predicting repetition does not seem to

consistently increase in the Analyst25-50-75 and Bingo25-50-75 Conditions, but there does seem to be a moderate increase in the Stock25-50-75 Condition.

Figure 11

Study 3B: Proportion of Participants Predicting Repetition of Terminal Streak, by Streak Length and Condition



Note: Results of Study 3B. Solid and patterned lines represent the proportion of participants in each Condition who predicted repetition of the streak at the end of each target sequence. Error bars represent +/– 1 standard error. In all three Conditions, the proportion of participants predicting repetition of the terminal streak increased between Streak Lengths 2 and 3. Across Streak Lengths 4 through 7, the proportion of participants predicting repetition does not seem to consistently increase in the Analyst25-50-75 or Bingo25-50-75 Conditions, but there does seem to be a moderate increase in the Stock25-50-75 Condition. N = 300 (Stock25-50-75 = 93; Analyst25-50-75 = 98; Bingo25-50-75 = 109).

Unlike Study 3A, we see no meaningful differences between predictions made by participants in the Analyst25-50-75 Condition and by those in the other two Conditions. Participants faced with an intentional actor were *not* more likely to predict repetition than those faced with a market or random mechanical generator.

Table 6

Study 3B: Proportion of Participants Predicting Repetition of Terminal Streaks, by Streak Length and Condition

		Proportion Who Predicted Streak Will Repeat					
Condition	N	Streak = 2	Streak = 3	Streak = 4	Streak = 5	Streak = 6	Streak = 7
Analyst25-50-75	98	0.27 (0.44)	0.42 (0.49)	0.53 (0.50)	0.50 (0.50)	0.60 (0.49)	0.55 (0.50)
Stock25-50-75	93	0.28 (0.45)	0.52 (0.50)	0.54 (0.50)	0.63 (0.48)	0.58 (0.49)	0.63 (0.48)
Bingo25-50-75	109	0.34 (0.47)	0.52 (0.50)	0.55 (0.50)	0.53 (0.50)	0.48 (0.50)	0.55 (0.50)
All	300	0.30 (0.46)	0.49 (0.50)	0.54 (0.50)	0.55 (0.50)	0.55 (0.50)	0.58 (0.50)

Note: Standard errors in parentheses.

Table 6 presents summary statistics for Study 3B, by Condition and Streak Length. We conducted a one-way mixed ANOVA to test the effects of one between-subjects variable (Condition), and one within-subjects variable (Streak Length) on participants' predictions that the streak at the end of each target sequence would *repeat*.[33] Participants' predictions ("1" for *repeat*; "0" for *reverse*) were the focus of analysis. The effect of Condition on participant predictions was not significant ($F(2, 297) = 0.62$, $p = 0.538$). And, Bonferroni-corrected pairwise comparisons revealed *no* significant differences between the proportion of participants who predicted streaks would repeat in the Bingo25-50-75 Condition and the proportion in the

[33] The results of a repeated measures binary logistic regression analysis of these data can be found in the *Binary Logistic Regression Analyses* section of the Online Supplement. There are no substantive differences between the results of the binary logistic analysis and those presented here.

Stock25-50-75 Condition (Mean Difference = –0.04, $p = 1.000$) or the Analyst25-50-75 Condition (Mean Difference = 0.02, $p = 1.000$). There was no significant difference between the proportion of participants predicting streaks would repeat in the Analyst25-50-75 and Stock25-50-75 Conditions either (Mean Difference = –0.05, $p = 0.816$).

There was a significant main effect of Streak Length on predictions, ($F(4.49, 1333.14) = 22.06$, $p < 0.000$).[34] A higher proportion of participants predicted streaks would repeat as Streak Length increased. However, Bonferroni-corrected pairwise comparisons revealed that the only significant differences were between Streak Length 2 and each of the other Streak Lengths. There were no significant differences between any combination of Streak Lengths 3–7. There was no significant interaction between Streak Length and Condition ($F(8.98, 1333.14) = 1.58$, $p = 0.115$). Participants responded to increases in Streak Length similarly across all three Conditions.

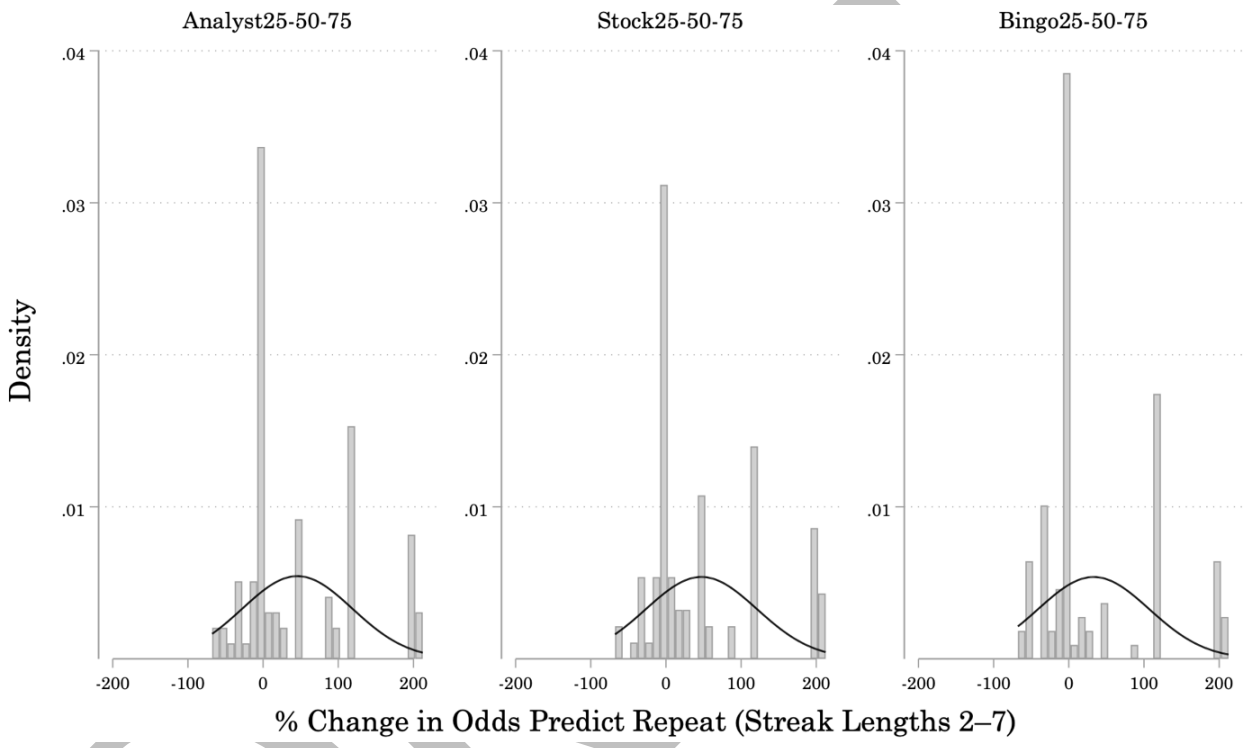
We again fitted logistic regressions to each participant's predictions for the six target sequences ending in a streak, and transformed the resulting coefficients from the log odds to the percent-change in the odds of predicting repetition for each unit increase in Streak Length.[35] Figure 12 presents the distribution of participant strategies in each Condition. The distributions of percent-change values are centered above 50% in the Analyst50 and Stock50 Conditions, and the distribution of values is centered above 25% in the Bingo50 Condition. The odds that a participant in the Bingo25-50-75 Condition predicts repetition for a streak of length x were about 25% higher than the odds a participant in this Condition predicts repetition for a streak of length

[34] Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(14) = 80.15$, $p < 0.000$); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.90$).
[35] Separation was observed while running participant-level logistic regressions on 92/300 participants' predictions over target sequences. Firth's procedure was applied to all of the participant-level logistic regressions to resolve the separation issue.

$x - 1$. A minority of participants exhibit negative percent-changes in the odds they will predict "repeat" as Streak Length increases: 18% in the Analyst25-50-75 Condition, 16% in the Stock25-50-75 Condition, and 28% in the Bingo25-50-75 Condition.

Figure 12

Study 3B: Distribution of Individual Participant's Prediction Strategies



Note: Individual participants' prediction strategies in Study 3B. Logistic regressions were fitted to participants' predictions over target stimuli ending in streaks of length 2–7 to find the percent-change in the odds a participant predicts a streak will repeat for each unit increase in the length of that streak. Histograms of frequencies for different values of percent-changes (gray bars), overlaid with normal density curves (dark gray lines). The distributions of percent-changes for participants in the Analyst25-50-75 and Stock25-50-75 Conditions are centered above 50%, and the distribution for participants in the Bingo25-50-75 Condition is centered above 25%.

Discussion: Studies 3A and 3B

Participants' predictions in Studies 3A and 3B are mostly consistent with rational Bayesian updating for all three generators. When provided with an explicit distribution of possible rates, participants seemed to use this information appropriately by updating their beliefs toward the maximum possible rate as Streak Length increased. Participants also made similar predictions about each of the three generators. The significant differences we saw between predictions made by participants facing the bingo cage and market generators in previous studies disappeared in the Bingo25-50-75 and Stock25-50-75 Conditions of Studies 3A and 3B. We saw a slight overreaction to streaks when participants were asked to provide a Continuous Response in the Analyst25-50-75 Condition of Study 3A, which we interpreted as weak evidence that participants may be more likely to exhibit a positive recency bias when evaluating outcomes produced by an intentional actor. However, this overreaction was not observed when participants were asked to respond in a binary format in the Analyst25-50-75 Condition of Study 3B. We therefore conclude that, taken together, the results of Studies 3A and 3B do not support the hypothesis that intentionality of the generator increases expectations that streaks will repeat (hot hand). Instead, the results support our favored hypothesis that the hot hand pattern arises from a reasonable updating process.

We dichotomized the predictions of participants in Study 3A in order to compare them to predictions made by participants in Study 3B. Welch's *t*-test for unequal variances indicated that a smaller proportion of participants in each Condition of Study 3B predicted streaks would repeat than in each Condition of Study 3A ($\text{Bingo}_{\text{STUDY3B}} - \text{Bingo}_{\text{STUDY3A}} = -0.26$, *s.e.* = 0.03, $p < 0.000$; $\text{Analyst}_{\text{STUDY3B}} - \text{Analyst}_{\text{STUDY3A}} = -0.37$, *s.e.* = 0.03, $p < 0.000$; $\text{Stock}_{\text{STUDY3B}} - \text{Stock}_{\text{STUDY3A}} = -0.19$, *s.e.* = 0.03, $p < 0.000$). These differences were significant at almost every

Streak Length. Again, it does not seem participants were more likely to exhibit a preference for streaks to repeat when asked to provide a Binary Response (Study 3B) than when asked to provide a Continuous Response (Study 3A). The opposite seems to be the case.

Section VI: General Discussion and Future Directions

We had three goals for the present studies. The first was to determine whether participants would exhibit different patterns of predictions when information about the base rate was held constant while the qualitative description of the generator varied. The second was to identify the effect of uncertainty about the base rate on participants' predictions. Third, we wanted to test whether participants' predictions would differ when asked to use numerical probability scale versus a dichotomous choice format. The present studies are the first to provide participants with consistent information about the base rates for three types of generators: a random mechanical device, an intentional actor, and a financial market. They are also the first to systematically manipulate participants' level of certainty over the base rate of the generator. We conclude that the dominant process underlying most predictions for sequences of binary outcomes is reasonable belief-updating about an uncertain base rate. We also identify a pocket of anomalous predictions for sequences generated by a random mechanical device with a strong prior belief about its base rate.

Generator Description and Base Rate Manipulations

Participants exhibit similar prediction patterns for all three types of generators when the base rate is uncertain. Participants gradually increased their expectations that streaks of identical outcomes would repeat as the length of those streaks increased. We observed this behavior in Studies 1A and 1B when the base rate was ambiguous (no information was provided), and in Studies 3A and 3B when we specified the distribution of possible base rates. In the latter case,

participants' predictions were close to the posterior probabilities calculated from a rational Bayesian updating model. In these four studies, we observed the same updating pattern for all three types of generators, including random mechanical devices. In contrast to results from previous studies where instructions did not precisely control base rate information across generators, we do not find that predictions for intentional actors are significantly different from those for financial markets or random mechanical devices.

Participants exhibited a different prediction pattern for random mechanical devices when they were provided with a specific, stationary base rate in Studies 2A and 2B. Predictions for the intentional actor and the market remained close to the .50 base rate for shorter streaks (2–4 outcomes), but increased to slightly above 50% for longer streaks (5–7 outcomes). This is consistent with Rabin's conjecture that the mere *plausibility* of variation in the underlying base rate can give rise to hot hand patterns in predictions (people seem to find it plausible for intentional actors and markets to exhibit shifting performance rates). Predictions for the random mechanical device were slightly below the .50 base rate across all streak lengths. This was surprising because the conventional version of the gambler's fallacy implies expectations for reversal *increase* with Streak Length, but predictions were flat. We also did not see the U-shaped pattern hypothesized by Rabin's model, whereby expectations for reversal initially increase, and then decrease as the observer is compelled to update her beliefs about the base rate when the pattern of outcomes discredits the original base rate belief.

How might we explain the unusual predictions we observed for random mechanical devices with a stationary .50 base rate? As noted in the introduction, some people are conditioned to produce this pattern through scholastic training in mathematics, that essentially teaches students the Law of Small Numbers. In a review of mathematics textbooks used in the

United States between 1957 and 2004, Jones (2004) notes that teachers are most often directed to introduce the concept of probability (and randomness) using one of the following pseudo-random devices: marbles in a jar, papers in a hat, cubic dice, coins, or spinners. The devices almost always have stationary, equiprobable base rates (e.g. 0.50 for each face of the coin). "Students therefore learn...that the purpose of drawing a random sample is to *ensure* representativeness in order to gain knowledge about the population from the sample" (Harradine, Batanero, & Rossman, 2011, p. 240, emphasis added). Thus, students learn to expect small samples of outcomes from equiprobable random mechanical devices will "represent" their population parameters (Stohl, 2005).

Describing the generator as a random mechanical device with a stationary base rate (Studies 2A and 2B) probably evoked these Law of Small Numbers beliefs, leading to the expectation that a streak of one outcome would "correct itself" and "balance out" even in small sample of outcomes. The developmental trend in predictions for sequences like coin tosses is consistent with this interpretation. Preschool children have reasonable intuitions about probability, and if anything, demonstrate a slight bias toward repetition of streaks (Bogartz, 1965; Chiesi & Primi, 2009; Craig & Meyers, 1963; Derks & Paclisanu, 1967; Estes, 1962; Fischbein, 1975; Fischbein & Schnarch, 1997). Reversal expectations and the gambler's fallacy pattern of predictions increase with age (Chiesi & Primi, 2009; Derks & Paclisanu, 1967).

Participants' verbal reports on their own prediction strategies also support this Law of Small Numbers interpretation (acknowledging that *post hoc* verbal reports are only suggestive, cf. Nisbett & Wilson, 1977). At the end of the experimental procedure, participants were asked: "What was your strategy for predicting what would happen next? What information did you use to make your prediction?" Participants' responses were classified into one of 8 categories: 1)

balancing outcomes; 2) guessing; 3) estimating a proportion or counting outcomes, including references to updating estimates; 4) momentum or increasing probability of one outcome over the other; 5) “following instructions" (often reported to justify an prediction of 50% in Studies 2A and 2B); 6) deciding which “type” of generator produced the sequence, particularly with reference to the distribution of rates provided in Studies 3A and 3B (e.g. high- versus low-performing analysts); 7) performing some sort of weighting calculation that takes into account the different types of generators, particularly in Studies 3A and 3B; 8) “other” unclassifiable responses. (See Appendix C: Verbal Reports for a summary of the methods and results of this “think-aloud” exercise.)

There was considerable variety in participants’ verbal reports in every experimental condition, but we see that references to estimating or updating proportions comprise more than 50% of the responses for all experimental conditions (categories labeled “Proportion” and “Momentum” in *Verbal Reports*). This fits our conclusion that base rate (proportion) updating is the primary inference process underlying the pervasive positive recency prediction patterns in our experiments. Second, “balancing” reports are scattered across experimental conditions and occur at highest rates among participants faced with sequences produced by a random mechanical device in the Bingo Conditions of each study (18% and 30% for the stationary .50 base rate in Studies 2A and 2B, respectively). This is consistent with our conclusion that Law of Small Numbers (“balancing outcomes”) reasoning is most likely to occur for random mechanical devices. Third, unsurprisingly, self-reports of reasoning about “types” of generators (coding categories labeled “Type” and “Weighting”) are common when “types” (e.g. Bad, Average, and Good analysts) are mentioned explicitly in the experimental instructions, as in Studies 3A and 3B.

As we noted in the introduction to this paper, another explanation for anomalous judgments of sequences generated by random mechanical devices is that observers transfer valid beliefs from analogous situations they have encountered outside the laboratory (Hahn & Warren, 2009; Kareev, 2000; Miller & Sanjurjo, 2018; and Reimers, Donkin, & Le Pelley, 2018, spell out detailed versions of this interpretation). This transfer process could be a simple generalization from the statistical properties of one situation to the new, to-be-judged situation (Turk-Browne, Scholl, Chun, & Johnson, 2009). Or the statistical regularities in one situation could be used to construct a mental model of a causal mechanism, and then that abstracted mechanism would be applied to deduce predictions in a new context. The problem with these transfer-of-statistical-patterns interpretations is that no one knows which extra-laboratory learning experiences observers actually rely on. Without that information it is not possible to provide a strong test of this highly plausible, but imprecise, family of interpretations. One might ask, for example, why the anomalous gambler's fallacy predictions occur at high rates only for random mechanical generators, when small samples and negative recency patterns occur for many other generators outside the laboratory.

Response Format

For the most part, predictions made by participants using a numerical probability scale are similar to those made by participants asked to make a dichotomous choice. However, if we interpret the *proportions* of participants *who predict repetitions* as degrees of belief, we find a couple of anomalies in the predictions made by participants responding with a dichotomous choice. First, the proportion of participants predicting repetition for streaks of two identical outcomes is surprisingly low (between .30 and .35) across all three studies. When given no information about the base rate, participants asked to make a dichotomous choice (Study 1B)

seem to update their beliefs at about the same rate as participants responding with numerical probability estimates (Study 1A). However, participants given a stationary base rate, or a specified distribution of possible rates, update more conservatively when asked to make a dichotomous choice (Studies 2B and 3B) than when asked to provide a numerical probability estimate (Studies 2A and 3A). This difference is especially pronounced for longer streaks, with the proportion of participants predicting repetition relatively flat across streak lengths 5–7 in Studies 2B and 3B.

Limitations

The present studies are subject to several limitations. First, we only presented participants with relatively short sequences of 8 outcomes, so we can't draw any strong conclusions about people's judgment behavior when exposed to longer sequences of outcomes. Second, despite our efforts to balance statistical properties across the full set of sequence stimuli, we still only used a small subset of possible sequences, and sequences with streaks of length 4 and greater were overrepresented (compared to a true random binomial process). Our research design was focused on systematically manipulating the lengths of terminal streaks at the end of an 8-event sequence. We did not systematically manipulate the proportion of each signal type in each of the target sequences. So, we cannot determine whether the proportion of each outcome in a given sequence influences predictions, independent of Streak Length. In hindsight, we consider this a weakness of our experimental plan.

Post hoc analyses that include our filler sequences (ending in reversals) show that *non-terminal* streaks and simple *global proportions* of outcomes across all eight events were associated with belief-updating prediction patterns, although the rate of updating was greater when the streak of similar events occurred at the end of a sequence. We conjecture that

observers are more attentive to patterned terminal streaks than to simple proportions or streaks that occur earlier in the sequence. Although, note again that our experimental filler stimulus sequences were not designed to provide a strong test of these relationships.

We also sacrificed control over the experimental task by recruiting all of our participants through MTurk. Participants may have been subject to distractions, and some likely manipulated the presentation of the experimental materials through browser-based scripts. It is also difficult to define the population represented by Workers on MTurk. Replications of our studies in a controlled laboratory environment are obviously necessary to validate our results.

Finally, the artificial design of our experimental procedure, though consistent with prior work, strips away potentially critical features of the information environment – for example, the emotionally-charged experience of watching a live basketball game, the thrill of a big payout from a successful bet, or the social and emotional consequences of a major loss on a stock market investment. It is possible that, without cues that evoke emotional or motivational responses in the observers, the "hot hand" or momentum phenomena is less likely to occur.

Concluding Remarks

The present research provides the most comprehensive overview of predictions in sequences of binary events in the scientific literature on human judgments. Our goal was to say something general about the conditions that produce the two most common prediction patterns, the hot hand and the gambler's fallacy. Our conclusion on this point is that the hot hand judgment pattern is most likely to occur after an observer sees a streak of similar outcomes. We submit that this pattern is best interpreted as the result of a reasonable (or even rational) bottom-up, evidence-based process for updating beliefs about the generator's base rate. We observe this pattern for all three types of generators when prior beliefs about the generator's base rate is

uncertain, regardless of whether participants express their predictions as numerical probability estimates or dichotomous choices.

However, when people hold strong prior beliefs about a stationary base rate, prediction patterns for a random mechanical device are different than those for an intentional actor or a market. Participants faced with a random mechanical device that has a stationary base rate exhibit a bias toward reversal of streaks. (Notably, the magnitude of this bias does not increase with Streak Length.) This bias toward reversal is stronger when participants make their predictions as dichotomous choices. This anomalous habit may result from mis-interpretations of principles of probability that participants learned in mathematics classes. Or, it may be that memories of non-laboratory experiences with sequences are being transferred to the controlled, focused experiences provided in our experiments. Our experiments were not designed to discriminate between these two accounts.

Notably, we found no compelling evidence for causal momentum or intentionality effects for any generators, beyond those implied by reasonable updating of beliefs about the generators' base rates.[36] People may expect streaks produced by intentional actors to repeat because it is true that intentional human performance often exhibits positively correlated sequences of outcomes. But, in the present research, belief updating provides a sufficient and more plausible explanation of positive recency prediction patterns. We also find that participants tend to expect streaks produced by a financial market to repeat. Here again, we think belief updating is the dominant cognitive process. We suspect the difference between our results and those of related experiments where participants exhibited a bias toward market streak reversal is due to small differences in experimental instructions. In contrast to many other studies' methods, our

[36] This is why the term "hot hand" does not appear in the title of this paper.

experiments provided relatively little information about the nature of the markets (or assets) that produced our sequences, and we did not allude to the random appearance of stock price sequences. And, we would note that there is great variety in expertise and personal investment theories associated with the varied results reported in other studies of stock market forecasting.

The present studies advance our understanding of the conditions under which two general judgment patterns, hot hand and gambler's fallacy, are likely to dominate individual judgments. Our emphasis on the pervasiveness of close-to-rational belief-updating prediction strategies and the critical moderating role of uncertainty about the base rate is new and important. This pattern appears clearly in many observers' predictions when judging a sequence generated by a random mechanical device that has a concrete, stationary base rate, especially when observers respond using a dichotomous choice format. But, we do not believe it is necessary to posit spooky causal beliefs, irrationally linking the outcomes produced by a random generator. Gambler's fallacy patterns of predictions seem most likely to derive from experiences with a subset of naturally-occurring sequences that actually exhibit negative recency and classroom experiences that essentially teaches students to believe in the Law of Small Numbers.

References

Alter, A. L., & Oppenheimer, D. M. (2006). From a fixation on sports to an exploration of mechanism: The past, present, and future of hot hand research. *Thinking and Reasoning, 12*, 431–444.

Altmann, E. M., & Burns, B. D. (2005). Streak biases in decision making: Data and a memory model. *Cognitive Systems Research, 6*(1), 5-16.

Amir, G. S., & Williams, J. S. (1999). Cultural influences on children's probabilistic thinking. *The Journal of Mathematical Behavior, 18*(1), 85-107.

Anderson, N. H. (1960). Effect of first-order conditional probability in a two-choice learning situation. *Journal of Experimental Psychology, 59*, 73-93.

Anderson, M. J., & Sunder, S. (1995). Professional traders as intuitive Bayesians. *Organizational Behavior and Human Decision Processes, 64*(2), 185-202.

Asparouhova, E., Hertzel, M., & Lemmon, M. (2009). Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers. *Management Science, 55*(1), 1766-1782.

Avugos, S., Bar-Eli, M., Ritov, I., & Sher, E. (2013a). The elusive reality of efficacy–performance cycles in basketball shooting: an analysis of players' performance under invariant conditions. *International Journal of Sport and Exercise Psychology, 11*(2), 184-202.

Ayton, P., Hunt, A. J., & Wright, G. (1989). Psychological conceptions of randomness. *Journal of Behavioral Decision Making*, *2*(4), 221-238.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness?. *Memory & Cognition*, *32*(8), 1369-1378.

Baquero, G., & Verbeek, M. (2015). Hedge fund flows and performance streaks: How investors weigh information (No. ESMT-15-01). *ESMT European School of Management and Technology*.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics, 49*, 307-343.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise, 7*(6), 525-553.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics, 12*, 428–454.

Barron, G., & Leider, S. (2010). The role of experience in the gambler's fallacy. *Journal of Behavioral Decision Making, 23*(1), 117-129.

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., & Sánchez, E. (2016). Research on teaching and learning probability. In *Research on teaching and learning probability* (pp. 1-33). Springer, Cham.

Blinder, D. S., & Oppenheimer, D. M. (2008). Beliefs about what types of mechanisms produce random sequences. *Journal of Behavioral Decision Making*, *21*(4), 414-427.

Bloomfield, R., & Hales, J. (2002). Predicting the next step of a random walk: experimental evidence of regime-shifting beliefs. *Journal of Financial Economics, 65*(3), 397-414.

Bogartz, R. S. (1965). Sequential dependencies in children's probability learning. *Journal of Experimental Psychology*, *70*(4), 365-370.

Boynton, D. M. (2003). Superstitious responding and frequency matching in the positive bias and gambler's fallacy effects. *Organizational Behavior and Human Decision Processes, 91*, 119 – 127.

Braga, J. N., Ferreira, M. B., Sherman, S. J., Mata, A., Jacinto, S., & Ferreira, M. (2018). What's next? Disentangling availability from representativeness using binary decision tasks. *Journal of Experimental Social Psychology, 76*, 307-319.

Bulkley, G., & Harris, R. D. F. (1997). Irrational analysts' expectations as a cause of excess volatility in stock prices. *Royal Economic Society, 107*(441), 359-371.

Burns, B. D. (2002). Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand." *Cognitive Psychology*, *48*(3), 295-331.

Burns, B. D. (2003). When it is adaptive to follow streaks: Variability and stocks. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 25, No. 25).

Burns, B. D., & Corpus, B. (2004). Randomness and inductions from streaks: "Gambler's fallacy" versus" hot hand." *Psychonomic Bulletin & Review*, *11*(1), 179-184.

Caruso, E. M., Waytz, A., & Epley, N. (2010). The intentional mind and the hot hand: Perceiving intentions makes streaks seem likely to continue. *Cognition*, *116*(1), 149-153.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree - An open-source platform for laboratory, online and field experiments. *Journal of Behavioral and Experimental Finance, 9*, 88-97.

Chiesi, F., & Primi, C. (2009). Recency effects in primary-age children and college students. *International Electronic Journal of Mathematics Education, 4*(3), 259-279.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conrad, J., & Kaul, G. (1998). An anatomy of trading strategies. *The Review of Financial Studies, 11*(3), 489-519.

Craig, G. J., & Meyers, J. L. (1963). A developmental study of sequential two-choice decision-making. *Child Development, 34*, 483-493.

Croson, R., & Sundali, J. (2005). The gambler's fallacy and the hot hand: Empirical data from casinos. *Journal of Risk and Uncertainty*, *30*(3), 195-209.

De Bondt, W. F. M. (1991). What do economists know about the stock market? *Journal of Portfolio Management, 17*(2), 84-91.

De Bondt, W. F. M. (1993). Betting on trends: Intuitive forecasts of financial risk and return. *International Journal of Forecasting, 9,* 355-371.

De Bondt, W. F., & Thaler, R. H. (1989). Anomalies: A mean-reverting walk down Wall Street. *Journal of Economic Perspectives, 3*(1), 189-202.

De Bondt, W. F. M., & Thaler, R. H. (1990). Do security analysts overreact? *The American Economic Review, 80*(2), 52-57.

Derks, P. L., & Paclisanu, M. I. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology, 73*(2), 278.

Diener, D., & Thompson, W.B. (1985). Recognizing randomness. *The American Journal of Psychology*, *98*(3), 433-447.

Dohmen, T., Falk, A., Huffman, D., Marklein, F., & Sunde, U. (2009). Biased probability judgment: Evidence of incidence and relationship to economic outcomes from a representative sample. *Journal of Economic Behavior & Organization*, *72*(3), 903-915.

Durham, G. R., Hertzel, M. G., & Martin, J. S. (2005). The market impact of trends and sequences in performance: New evidence. *The Journal of Finance, 60*(5), 2551-2569.

Edwards, W. (1961). Probability learning in 1,000 trials. *Journal of Experimental Psychology, 62*, pp. 381-390.

Estes, W. K. (1962). Learning theory. *Annual Review of Psychology, 13*, 107-144.

Estes, W. K. (1964). Probability learning. In A.W. Melton (Ed.), *Categories of Human Learning* (pp. 88–128). New York: Academic Press.

Falk, R. (1981). The perception of randomness. In *Proceedings of the Fifth International Conference for the Psychology of Mathematics Education* (pp. 222-229). Grenoble, France.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications* (3rd Edition, Volume 1). New York: John Wiley.

Fiorina, M.P. (1971). A note on probability matching and rational choice. *Behavioral Science, 16(2),* 158-166.

Fischer, I., & Savranevski, L. (2015). Extending the two faces of subjective randomness: From the gambler's and hot-hand fallacies toward a hierarchy of binary sequence perception. *Memory & cognition*, *43*(7), 1056-1070.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht, The Netherlands: Reidel.

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*, 96-105.

Forbes, W. P. (1995). Picking winners? A survey of the mean reversion and overreaction of stock prices literature. *Journal of Economic Surveys, 10*(2), 123-158.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295-314.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499-510.

Gronchi, G., & Sloman, S. A. (2008). Do causal beliefs influence the hot-hand and the gambler's fallacy?. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1164-1168). Cognitive Science Society. Austin, TX.

Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: why three heads are better than four. *Psychological Review, 116*(2), 454.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on systems, man, and cybernetics, 17*(5), 753-770.

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 235-246). Springer, Dordrecht.

Harris, R. J. (1985). *A primer of multivariate statistics* (2nd ed.). New York: Academic Press.

Hawkins, A. S., & Kapadia, R. (1984). Children's conceptions of probability—a psychological and pedagogical review. *Educational Studies in Mathematics, 15*(4), 349-377.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine, 21*(16), 2409–2419.

Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *Journal of Experimental Psychology, 41*(4), 291–297.

Jegadeesh, N., & Titman, S. (2011). Momentum. *Annual Review of Financial Economics, 3*(1), 493-509.

Johnson, J., Tellis, G. J., & Macinnis, D. J. (2005). Losers, winners, and biased trades. *Journal of Consumer Research, 32*, 324-329.

Jones, D. L. (2004). *Probability in middle grades mathematics textbooks: An examination of historical trends, 1957–2004*. University of Missouri-Columbia.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107*(2), 397.

Koehler, J. J., & Conley, C. A. (2003). The "hot hand" myth in professional basketball. *Journal of Sport and Exercise Psychology*, *25*(2), 253-259.

Konold, C. (1995). Confessions of a coin flipper and would-be instructor. *The American Statistician, 49*(2), 203-209.

Laplace, P. S. (1902). *A philosophical essay on probabilities* (F. W. Truscott & F. L. Emory, Translators). New York, NY: John Wiley & Sons. (Original work published 1814).

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557-568.

Lee, W. (1971). *Decision Theory and Human Behavior* (pp. 163-167). New York: John Wiley & Sons.

Lindman, H., & Edwards, W. (1961). Supplementary report: Unlearning the gambler's fallacy. *Journal of Experimental Psychology, 62*(6), 630.

Loh, R. K., & Warachka, M. (2012). Streaks in earnings surprises and the cross-section of stock returns. *Management Science, 58*(7), 1305-1321.

Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8*, 626-636.

Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13*(3), 392-400.

Militana, E., Wolfson, E., & Cleaveland, J. M. (2010). An effect of inter-trial duration on the gambler's fallacy choice bias. *Behavioural Processes, 84*(1), 455-459.

Miller, J. B., & Sanjurjo, A. (2018a). How experience confirms the gambler's fallacy when sample size is neglected. OSF Preprints, 30.

Miller, J. B., & Sanjurjo, A. (2018b). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, *86*(6), 2019-2047.

Morrison, R.S., & Ordeshook, P.C. (1975). Rational choice, light guessing and the gambler's fallacy. *Public Choice, 22(1),* 79-89.

Morsanyi, K., Handley, S. J., & Serpell, S. (2013). Making heads or tails of probability: An experiment with random generators. *British Journal of Educational Psychology, 83*, 379-395.

Murphy, G.L., & Ross, B.H. (2010). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(2),* 263-276.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109*, 330–357.

Nicks, D. C. (1959). Prediction of sequential two-choice decisions from event runs. *Journal of Experimental Psychology, 57*(2), 105.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84(3),* 231-259.

Olivola, C. Y., & Oppenheimer, D. M. (2008). Randomness in retrospect: Exploring the interactions between memory and randomness cognition. *Psychonomic Bulletin & Review, 15*(5), 991-996.

Önkal, D., & Muradoglu, G. (1996). Effects of task format on probabilistic forecasting of stock prices. *International Journal of Forecasting, 12*(1), 9-24.

Oskarsson, A. T., Boven, L. V., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin, 135*, 262–285.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, *117*(3), 775-816.

Rabin, M., & Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, *77*(2), 730-778.

Rapoport, A., & Budescu, D. V. (1997). Randomization in individual choice behavior. *Psychological Review, 104*(3), 603.

Rao, J. M. (2009). Experts' perceptions of autocorrelation: The hot hand fallacy among professional basketball players. Unpublished technical manuscript. California. San Diego. Downloaded from http://www.justinmrao.com/playersbeliefs.pdf (July 11th, 2012).

Reimers, S., Donkin, C., Le Pelley, M. E. (2018). Perceptions of randomness in binary sequences: Normative, heuristic, or both? *Cognition, 172,* 11-25.

Restle, F. (1961). *The Psychology of Judgment and Choice: A Theoretical Essay.* New York: Wiley & Sons.

Restle, F. (1966). Run structure and probability learning: Disproof of Restle's model. *Journal of Experimental Psychology, 72*(3), 382-389.

Roney, C. J., & Trick, L. M. (2009). Sympathetic magic and perceptions of randomness: The hot hand versus the gambler's fallacy. *Thinking & reasoning, 15*(2), 197-210.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook o f research on mathematics teaching and learning* (pp. 465-494). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M., Canada, D., & Ciancetta, M. (2003). Middle school students' thinking about variability in repeated trials: A cross-task comparison. In N. A. Pateman, B. J. Dougherty & J. T. Zilliox (Eds.), *Proceedings o f the 2003 Joint Meeting of PME and PMENA* (Vol. 4, pp. 159-165). Honolulu, HI: University of Hawai'i.

Scholl, S., & Greifeneder, R. (2011). Disentangling the effects of alternation rate and maximum run length on judgments of randomness. *Judgment and Decision Making, 6*(6), 531-541.

Shanthikumar, D. M. (2012). Consecutive earnings surprises: Small and large trader reactions. *The Accounting Review, 87*(5), 1709-1736.

Soetens, E., Boer, L. C., & Hueting, J. E. (1985). Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 11*(5), 598.

Steinbring, H. (1990). The use of chance-concept in everyday teaching – Aspects of socially constituted epistemology of mathematical knowledge. In J. B. Garfield (Ed.), *Research papers from the Third International Conference on Teaching Statistics*. University of Otago, Dunedin, New Zealand.

Stohl, H. (2005). Probability in teacher education and development. In *Exploring probability in school* (pp. 345-366). Springer, Boston, MA.

Suetens, S., Galbo-Jorgensen, C.B., & Tyran, J.-R. (2016). Predicting Lotto numbers: A natural experiment on the Gambler's Fallacy and the Hot-Hand Fallacy. *Journal of the European Economic Association, 14(3),* 584-607.

Sun, Y., & Wang, H. (2010). Gambler's fallacy, hot hand belief, and the time of patterns. *Judgment and Decision Making, 5*(2), 124-132.

Turk-Browne, N.B., Scholl, B.J., Chun, M.M., & Johnson, M.K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience, 21(10),* 1934-1945.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105.

Tyszka, T., Markiewicz, Ł., Kubińska, E., Gawryluk, K., & Zielonka, P. (2017). A belief in trend reversal requires access to cognitive resources. *Journal of Cognitive Psychology, 29*(2), 202-216.

Tyszka, T., Zielonka, P., Dacey, R., & Sawicki, P. (2008). Perception of randomness and predicting uncertain events. *Thinking & Reasoning, 14*(1), 83-110.

Vergin, R. C. (2000). Winning streaks in sports and the misperception of momentum. *Journal of Sport Behavior, 23*(2), 181.

Windschitl, P.D., & Wells, G.L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: General, 2(4),* 343-364.

Appendix A: Stimuli

Description of stimuli sequences

The same stimuli sequences were used in all of the reported experiments. Though the images representing the outcomes in each sequence differed across Conditions in each experiment, the patterns of outcomes were the same across Conditions. There are two pools of stimuli: Targets and Fillers. The Target stimuli pool contains 22 sequences, each comprised of 8 signals, and each ending in a streak of at least two identical signals. For each terminal streak length 2 through 6, two different sequences were created. Two sequences end in a streak of Red/Up outcomes, and two sequences end in a streak of Blue/Down outcomes. The two sequences ending in a streak of Red/Up outcomes are mirror images of the two sequences ending in a streak of Blue/Down outcomes. For the terminal streak length 7, two sequences were created. One ends in 7 Red/Up outcomes, and one ends in 7 Blue/Down outcomes.

The patterns of outcomes that precede the terminal streak in each Target sequence were designed to achieve the several goals: 1) achieve an average alternation rate close to 0.50 across all Target sequences, 2) include sequences having the same terminal streak length, but different alternation rates, 3) include sequences having the same terminal streak length, but different Bayesian posterior probabilities. The third goal relates only to Studies 3A and 3B, where a distribution of possible rates was provided to participants, which made it possible to estimate the posterior probability that each sequence's terminal streak would repeat, conditional on the pattern of outcomes in that sequence, and the distribution of possible rates. The full list of Target stimuli sequences is enumerated in Table A1.

Table A1

Description of Target Stimuli Sequences.

Sequence	Terminal Streak Type	Terminal Streak Length	Proportion Streak Type	Alternation Rate	Bayesian Posterior
01001011	1	2	0.50	0.71	0.50
10110100	0	2	0.50	0.71	0.50
01010100	0	2	0.63	0.86	0.60
10101011	1	2	0.63	0.86	0.60
01010111	1	3	0.63	0.71	0.60
10101000	0	3	0.63	0.71	0.60
01011000	0	3	0.63	0.57	0.60
10100111	1	3	0.63	0.57	0.60
00101111	1	4	0.63	0.43	0.60
11010000	0	4	0.63	0.43	0.60
01010000	0	4	0.75	0.57	0.68
10101111	1	4	0.75	0.57	0.68
00100000	0	5	0.88	0.29	0.72
11011111	1	5	0.88	0.29	0.72
01011111	1	5	0.75	0.43	0.68
10100000	0	5	0.75	0.43	0.68
01000000	0	6	0.88	0.29	0.72
10111111	1	6	0.88	0.29	0.72
11000000	0	6	0.75	0.14	0.68
00111111	1	6	0.75	0.14	0.68
01111111	1	7	0.88	0.14	0.72
10000000	0	7	0.88	0.14	0.72
Mean			**0.72**	**0.47**	**0.65**

Note: Terminal Streak Type is the signal type repeated in the terminal streak at the end of each sequence (0 or 1). Terminal Streak Length is the number of identical signals repeated at the end of each Target sequence. Proportion Streak Type is the proportion of signals in a given sequence that match the signal type repeated in the terminal streak. Bayesian Posterior applies to Study 3A and 3B only. It is the posterior probability that a given terminal streak will repeat, conditional on the pattern of signals in the sequence, and the distribution of probability rates provided in Study 3, $p(0.25) = p(0.50) = p(0.75) = 0.33$. (Coding: 1 = Up/Red, 0 = Down/Blue.)

Table A2

Description of Filler Stimuli Sequences

Sequence	Longest Streak Type	Longest Streak Length	Proportion Streak Type	Alternation Rate	Bayesian Posterior
01010101	–	–	–	1.00	0.50
10101010	–	–	–	1.00	0.50
00110101	1	2	0.50	0.71	0.50
00110110	1	2	0.50	0.57	0.50
01010010	0	2	0.63	0.86	0.60
01101010	1	2	0.50	0.86	0.50
10100110	1	2	0.50	0.71	0.50
11001101	1	2	0.63	0.57	0.60
11010010	0	2	0.50	0.71	0.50
11011001	0	2	0.38	0.57	0.60
01000101	0	3	0.63	0.71	0.40
01011101	1	3	0.63	0.71	0.60
01110010	1	3	0.50	0.57	0.50
10001010	0	3	0.63	0.71	0.60
11000110	0	3	0.50	0.43	0.50
11101010	1	3	0.63	0.71	0.40
00001101	0	4	0.63	0.43	0.40
01011110	1	4	0.63	0.57	0.40
10000101	0	4	0.63	0.57	0.40
11110010	1	4	0.63	0.43	0.40
01000001	0	5	0.75	0.43	0.32
11111010	1	5	0.75	0.43	0.32
10000001	0	6	0.75	0.29	0.32
11111101	1	6	0.88	0.29	0.72
Mean			**0.59**	**0.62**	**0.48**

Note: Longest Streak Type is the signal type repeated in the longest streak in each sequence (0 or 1). (For sequences with multiple "longest streaks" of length 2, we take the signal type of streak closest to the end of the sequence.) Longest Streak Length is the number of identical signals repeated in the longest streak in each sequence. Proportion Streak Type is the proportion of signals in a given sequence that match the signal type repeated in the terminal streak. Bayesian Posterior applies to Study 3A and 3B only. It is the posterior probability that the final (8th) signal in the sequence will repeat, conditional on the pattern of signals in the sequence, and the distribution of probability rates provided in Study 3, $p(0.25) = p(0.50) = p(0.75) = 0.33$. (Coding: 1 = Up/Red, 0 = Down/Blue.)

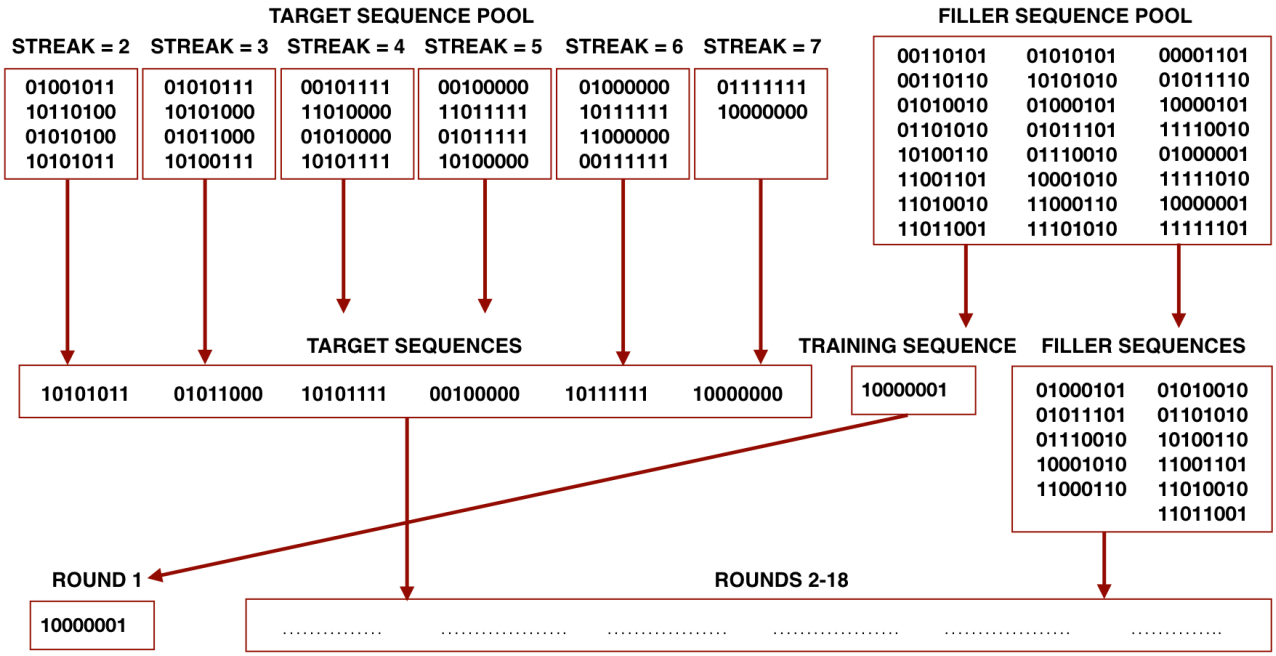
The Filler stimuli pool contains 24 sequences, each comprised of 8 signals, and each ending in a reversal (e.g. Red-Blue, Up-Down). The patterns of outcomes in each Filler sequence were designed to achieve several goals: 1) achieve an average alternation rate close to 0.50 across both Target and Filler sequences, 2) introduce streaks of various lengths that appear toward the beginning and middle of a given sequence, 3) achieve an average Bayesian posterior probability close to 0.50 across both Target and Filler sequences, 4) balance the total number of Red/Up signals and Blue/Down signals across all Target and Filler sequences. The full list of Filler stimuli sequences is enumerated in Table A2.

Method of Randomization

The same procedure for randomly selecting, and presenting, stimuli sequences was used in all of the reported experiments. Each participant was shown 18 sequences – 12 Filler sequences and 6 Target sequences. For each participant, the 12 Filler sequences were randomly drawn from a pool of 24 sequences ending in reversals. The 6 focal stimuli were randomly selected from the pool of 22 sequences ending in streaks. One sequence was selected for each Streak Length 2 through 7. We prepared 4 Target stimuli ending in each Streak Length 2 through 6, and 2 stimuli ending in a Streak of Length 7. For each Streak Length, half of the stimuli end in Red/Up streaks, and half end in Blue/Down streaks. In this way, we were able to alternate the signal type of the terminal streak, and introduce variation in the patterns preceding the terminal streak in each target sequence. One of the 12 Filler sequences is randomly selected to appear in the first round. The remaining 11 Filler sequences, and 6 Target sequences, are shuffled and presented in random order across rounds 2-18. Figure A1 illustrates the randomization process.

Figure A1

Illustration of Process Used to Randomly Select and Present Stimuli to Participants



Appendix B: Procedure

Instructions

The transcripts of the instructions provided to participants in each of the present studies can be found below. Instructions were identical across "A" and "B" versions of each study.

Studies 1A and 1B – Unknown (Ambiguous) Rate

> **[AnalystUnknown Condition]**
> Stock analysts look for trends in the stock market, and use that information to invest their clients' money wisely. Most stock analysts manage a "book" of stocks, which is a collection of investments they've made for their clients. Each quarter, analysts will evaluate trends in the market, and decide whether to sell some of the stocks in their book, purchase more of those stocks, or purchase new (different) stocks. **The result of these decisions determines whether the total value of their book increases or decreases in the subsequent quarter.**
>
> In this task, you will see the **outcomes of investment decisions made by several stock analysts**. We will reveal to you, one at a time, the **change in the value of each analyst's book over the course of eight quarters** (2 years). If an analyst made good decisions in the prior quarter, then you will see the value of that analyst's book go up, if the analyst made poor decisions, you will see the value of that analyst's book go down. **After watching what happens to each analyst's book** over the course of eight quarters, **your job is to predict what will happen to the analyst's book in the ninth (next) quarter**: Do you predict it will go up or down? After each prediction, we will move on and show you the results for a new, different stock analyst. This process will repeat 18 times, so you'll see 18 different stock analysts' performance histories in total. **Remember that each round** you will see the changes in value of one analyst's book over 8 quarters, and you will make a prediction about how the value of that analyst's book will change in the 9th (next) quarter.
>
> **[StockUnknown Condition]**
> Stock prices change constantly. **Price movement of a stock reflects the market's evaluation, and buyers' and sellers' expectations of a company's worth**. Many factors influence stock prices, such as earnings reports, news about a company's leadership and products, economic policies, and political events.
>
> In this task, you will observe **changes in different companies' stock prices.** We will reveal to you, one at a time, the **change in the value of each company's stock price over the course of eight quarters** (2 years). If the market evaluates a company's worth as higher than it was worth the previous quarter, then you will see that company's stock price go up. If the market evaluates a company's worth as lower than it was worth the previous quarter, you will see that company's stock price go down. **After watching what happens to each company's stock price** over the course of eight quarters, **your**

job is to predict what will happen to the company's stock price in the ninth (next) quarter: Do you predict it will go up or down? After each prediction, we will move on and show you the results for a new, different company. This process will repeat 18 times, so you'll see 18 different companies' stock price histories in total. **Remember that each round** you will see the changes in one company's stock price over 8 quarters, and you will make a prediction about how that company's stock price will change in the 9th (next) quarter.

[BingoUnknown Condition]
In this task you will watch a mechanical bingo machine draw red and blue balls from its **covered** cage. The cage contains **a mix of red and blue balls**. Since the cage is covered, **no one knows exactly how many red balls or how many blue balls are in the cage**. At the start of every round, the machine spins the cage, mixing up all of the balls, and then rolls one ball out of the cage so that an announcer sitting next to the cage can see the color of the ball. The announcer calls out the color of the ball ("red" or "blue"), and the machine rolls the ball back into its covered cage. The machine then spins the cage again, and rolls another ball out. The announcer calls out the color of the ball, and the machine rolls that ball back into its cage. The machine continues **randomly drawing colored balls from the cage until 9 balls have been drawn**. After the 9th ball has been drawn, the round ends, and there is a pause.

Each round, we will reveal to you, one at a time, the colors of the first 8 balls drawn by the machine. **After watching the outcomes** of these eight draws, **your job is to predict the color of the next (9th) ball drawn by the machine**. After you make your prediction about the 9th ball, you will move on to the next round, where you will watch the machine draw another 8 balls and you will again predict the color of the 9th ball. This process will repeat 18 times, so you will watch 18 different rounds of bingo ball draws. **Remember that each round** you will see the color of the balls from the first 8 draws, and you will make a prediction about the color of the 9th ball drawn.

Studies 2A and 2B – Stationary 50% Rate

[Analyst50 Condition]
Stock analysts look for trends in the stock market, and use that information to invest their clients' money wisely. Most stock analysts manage a "book" of stocks, which is a collection of investments they've made for their clients. Each quarter, analysts will evaluate trends in the market, and decide whether to sell some of the stocks in their book, purchase more of those stocks, or purchase new (different) stocks. **The result of these decisions determines whether the total value of their book increases or decreases in the subsequent quarter**. If an analyst has a successful quarter and the value of his or her book increases, he or she gets a bonus; but if the book drops in value, there is no bonus and sometimes the analyst has to pay a penalty fine.

In this task, you will see the **outcomes of investment decisions made by several stock analysts**. The analysts you will see all have the same level of skill. **All of the analysts are of Average skill level. The probability that each analyst's book of business will increase in value is always 50%**.

All of the analysts you will see have stable skill levels, and their **skill levels do not change over time**. We will reveal to you, one at a time, the **change in the value of each analyst's book over the course of eight quarters** (2 years). If an analyst made good decisions in the prior quarter, then you will see the value of the book go up, if the analyst made poor decisions, you will see the value of the book go down. **After watching what happens** to each analyst's book over the course of 8 quarters, **your job is to predict what will happen to that analyst's book in the next (9th) quarter**: Do you predict it will go up or down? After each prediction, we will move on and show you the results for a new, different stock analyst. This process will repeat 18 times, so you'll see 18 different stock analysts' performance histories in total. **Remember that each round** you see the changes in value of one analyst's book over 8 quarters, and you will make a prediction about how the value of that analyst's book will change in the 9th (next) quarter.

[Stock50 Condition]
Stock prices change constantly. **Price movement of a stock reflects the market's evaluation, and buyers' and sellers' expectations of a company's worth**. Many factors influence stock prices, such as earnings reports, news about a company's leadership and products, economic policies, and political events.

In this task, you will observe **changes in different companies' stock prices**. The companies you will observe all have the same level of performance. **All of the companies are Average performers. The probability that each company's stock price will increase in value is always 50%**.

All of the companies you will see have stable performance levels, and their **performance levels do not change over time**. We will reveal to you, one at a time, the **change in the value of each company's stock price over the course of eight quarters** (2 years). If the market evaluates a company's worth as higher than it was worth the previous quarter, then you will see the stock price go up. If the market evaluates a company's worth as lower than it was worth the previous quarter, you will see the stock price go down. **After watching what happens** to a company's stock price over the course of 8 quarters, **your job is to predict what will happen to that company's stock price in the next (9th) quarter**: Do you predict it will go up or down? After each prediction, we will move on and show you the results for a new, different company. This process will repeat 18 times, so you'll see 18 different companies' stock price histories in total. **Remember that each round** you will see the changes in one company's stock price over 8 quarters, and you will make a prediction about how that company's stock price will change in the 9th (next) quarter.

[Bingo50 Condition]
In this task you will watch a mechanical bingo machine draw red and blue balls from its **covered** cage. The cage contains **50 red balls and 50 blue balls**. At the start of every round, the machine spins the cage, mixing up all of the balls, and then rolls one ball out of the cage so that an announcer sitting next to the cage can see the color of the ball. The announcer calls out the color of the ball ("red" or "blue"), and the machine rolls the ball back into its covered cage. The machine then spins the cage again, and rolls another ball

out. The announcer calls out the color of the ball, and the machine rolls that ball back into its cage. The machine continues **randomly drawing colored balls from the cage until 9 balls have been drawn**. After the 9th ball has been drawn, the round ends, and there is a pause.

Each round, we will reveal to you, one at a time, the colors of the first 8 balls drawn by the machine. **After watching the outcomes** of these eight draws, **your job is to predict the color of the next (9th) ball drawn by the machine**. After you make your prediction about the 9th ball, you will move on to the next round, where you will watch the machine draw another 8 balls and you will again predict the color of the 9th ball. This process will repeat 18 times, so you will watch 18 different rounds of bingo ball draws. **Remember that each round** you will see the color of the balls from the first 8 draws, and you will make a prediction about the color of the 9th ball drawn.

Studies 3A and 3B – Specified Distribution of Rates (.25, .50, .75)

[Analyst25-50-75 Condition]
Stock analysts look for trends in the stock market, and use that information to invest their clients' money wisely. Most stock analysts manage a "book" of stocks, which is a collection of investments they've made for their clients. Each quarter, analysts will evaluate trends in the market, and decide whether to sell some of the stocks in their book, purchase more of those stocks, or purchase new (different) stocks. **The result of these decisions determines whether the total value of their book increases or decreases in the subsequent quarter**. If an analyst has a successful quarter and the value of his or her book increases, he or she gets a bonus; but if the book drops in value, there is no bonus and sometimes the analyst has to pay a penalty fine.

In this task, you will see the **outcomes of investment decisions made by several stock analysts**. The analysts you will observe have **different skill levels: Bad, Average, and Good**.

- Bad analysts' book values go up about 25% of the time (1 out of 4 quarters)
- Average analysts' book values go up about 50% of the time (2 out of 4 quarters)
- Good analysts' book values go up about 75% of the time (3 out of 4 quarters)

All of the analysts you will see have stable skill levels, and their **skill levels do not change over time**. There are about the same number of Bad, Average, and Good analysts. We will reveal to you, one at a time, the **change in the value of each analyst's book over the course of eight quarters** (2 years). If an analyst made good decisions in the prior quarter, then you will see the value of the book go up, if the analyst made poor decisions, you will see the value of the book go down. **After watching what happens** to each analyst's book over the course of 8 quarters, **your job is to predict what will happen to that analyst's book in the next (9th) quarter**: Do you predict it will go up or down? After each prediction, we will move on and show you the results for a new, different stock analyst. This process will repeat 18 times, so you'll see 18 different stock analysts' performance histories in total. **Remember that each round** you see the changes

in value of one analyst's book over 8 quarters, and you will make a prediction about how the value of that analyst's book will change in the 9th (next) quarter.

[Stock25-50-75 Condition]
Stock prices change constantly. **Price movement of a stock reflects the market's evaluation, and buyers' and sellers' expectations of a company's worth**. Many factors influence stock prices, such as earnings reports, news about a company's leadership and products, economic policies, and political events.

In this task, you will observe **changes in different companies' stock prices**. The companies you will observe have **different performance levels: Bad, Average, and Good**.

- Bad companies' stock prices go up about 25% of the time (1 out of 4 quarters)
- Average companies' stock prices go up about 50% of the time (2 out of 4 quarters)
- Good companies' stock prices go up about 75% of the time (3 out of 4 quarters)

All of the companies you will see have stable performance levels, and their **performance levels do not change over time**. There are about the same number of Bad, Average, and Good companies. We will reveal to you, one at a time, the **change in the value of each company's stock price over the course of eight quarters** (2 years). If the market evaluates a company's worth as higher than it was worth the previous quarter, then you will see the stock price go up. If the market evaluates a company's worth as lower than it was worth the previous quarter, you will see the stock price go down. **After watching what happens** to a company's stock price over the course of 8 quarters, **your job is to predict what will happen to that company's stock price in the next (9th) quarter**: Do you predict it will go up or down? After each prediction, we will move on and show you the results for a new, different company. This process will repeat 18 times, so you'll see 18 different companies' stock price histories in total. **Remember that each round** you will see the changes in one company's stock price over 8 quarters, and you will make a prediction about how that company's stock price will change in the 9th (next) quarter.

[Bingo25-50-75 Condition]
In this task you will watch a mechanical bingo machine draw red and blue balls from one of three **covered** cages. Each covered bingo cage contains a mix of 100 red and blue balls.

- Cage #1 contains 25 red balls and 75 blue balls.
- Cage #2 contains 50 red balls and 50 blue balls.
- Cage #3 contains 75 red balls and 25 blue balls.

At the start of every round, the machine **randomly selects ONE of the 3 cages**. Each cage has an equal probability of being selected. **Since the cages are covered, no one knows what the mix of red and blue balls is in the cage the machine selects**. The machine spins whichever cage it randomly selected, then rolls one ball out of that cage so

> that an announcer sitting next to the machine can see the color of the ball. The announcer calls out the color of the ball ("red" or "blue"), and the machine rolls the ball back into the cage. The machine continues **randomly drawing colored balls from the current cage until 9 balls have been drawn**. After the 9th ball has been drawn, there is a pause. **During that pause, the machine selects which covered bingo cage** to use **for the next 9 draws**. The machine is **equally likely to choose any cage**. Sometimes it picks the same cage from the prior round again, sometimes it chooses a new cage. **The machine's selection is completely random, and totally unrelated to the mix of red and blue balls in the cage it picks**.
>
> Each round, we will reveal to you, one at a time, the colors of the first 8 balls drawn by the machine. **After watching the outcomes** of these eight draws, **your job is to predict the color of the next (9th) ball drawn by the machine from the current cage**. After you make your prediction about the 9th ball, you will move on to the next round, where you will watch the machine draw another 8 balls and you will again predict the color of the 9th ball the machine will draw that round. This process will repeat 18 times, so you will watch 18 different rounds of bingo ball draws. **Remember that each round** you will see the color of the balls from the first 8 draws, and you will make a prediction about the color of the 9th ball drawn.

Comprehension Check Questions

In each of the present studies, participants were required to pass a comprehension check after reading the instructions. Participants were allowed to attempt the comprehension check questions as many times as they wished. In Study 1A, participants were not allowed to review the instructions between attempts. We observed a small amount of differential attrition across Conditions at the comprehension check stage of the Study 1A. 12 participants abandoned the procedure after reaching the StockUnknown comprehension check, versus 8 who abandoned after reaching the AnalystUnknown comprehension check, and 2 after reaching the BingoUnknown comprehension check. No other patterns of differential attrition appeared at any other stage of the procedure. In Studies 1B, 2A, 2B, 3A, and 3B the comprehension check page was updated so that the instructions appeared below the comprehension questions. Participants in these studies could refer to the instructions if they were struggling to answer any of the questions. We did not observe differential attrition across conditions in any of these studies.

The comprehension check questions tested participants understanding of the following:

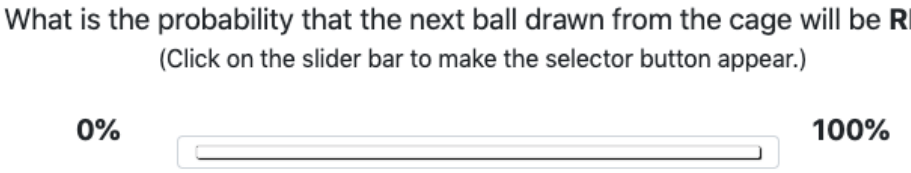
1. That their task was to predict the *next* outcome in the sequence
2. That each sequence of 8 outcomes was *new* and *not a continuation* of the previous sequences (in the Stock and Analyst Conditions, this meant that each sequence was produced by a *different* company or analyst, respectively).
3. That each sequence of 8 outcomes was produced consecutively by the *same* agent (bingo machine, stock analyst, company).
4. [Bingo Conditions Only] That draws from the bingo cage(s) were made *with replacement*.
5. [Studies 3A and 3B Only] That the rate at which each agent (bingo machine, stock analyst, company) produced Red/Up outcomes was either 25%, 50%, or 75%, and that there was an equal probability that a given agent produced Red/Up outcomes at each of these rates.

Prediction Prompts

Participants were asked to predict the next (9th) outcome of each sequence after watching the sequential revelation of the preceding 8 outcomes. Participants in Studies 1A, 2A, and 3A made their predictions using a continuous sliding scale labeled *0%* on the left-hand side, and *100%* on the right-hand side. Figure B1 presents a screenshot of the prediction prompt and scale used in the Bingo Conditions of these studies.[37] (The inputs for the Analyst and Stock Conditions were identical except for the wording of the question prompts.)

Figure B1

Screenshot of Prediction Prompt and Response Scale Used in The Bingo Conditions



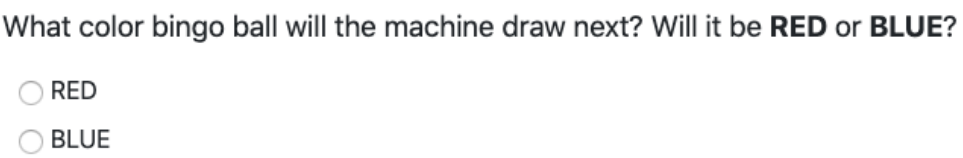
Participants in Studies 1B, 2B, and 3B made their predictions by selecting one of the two possible outcomes. Figure B2 presents a screenshot of the prediction prompt and radial response

[37] The slider selector button was hidden, so participants had to click on the slider range to make it appear. This precaution was taken to prevent participants from becoming anchored to the selector button's point of origin.

buttons used in the Bingo Conditions of these studies.[38] (The inputs for the Analyst and Stock Conditions were identical except for the wording of the question prompts.)

Figure B2

Screenshot of Prediction Prompt and Radial Response Used in Bingo Conditions



In the Bingo Conditions of Studies 1B, 2B, and 3B, participants were asked, "What color bingo ball will the machine draw next? Will it be **RED** or **BLUE**?" In the Analyst Conditions, participants were asked, "What will happen to the value of this analyst's book next quarter, will it go **UP** or **DOWN**?" In the Stock Conditions, participants were asked, "What will happen to the value of this company's stock next quarter? Will it go **UP** or **DOWN**?"

Recall that the dependent variable in the present studies is the participant's prediction that the last (8th) outcome in each sequence will *repeat*. However, participants were not asked to predict repetition directly; rather, participants in the continuous response ("A") version of each study were always asked for the probability that the next outcome would be Red (Bingo Conditions) or Up (Analyst and Stock Conditions), and participants in the binary choice ("B") versions of each study were always asked to choose which outcome they thought would occur next (Red/Up or Blue/Down). Because each target sequence was randomly drawn from several possible versions ending in either Red (Up) streaks or Blue (Down) streaks, each of the target sequences seen by a participant could have ended in a streak of Red (Up) signals or a streak of Blue (Down) signals.

[38] Red/Up always appeared as the top radial, and Blue/Down always appeared as the bottom radial.

For this reason, participants' responses needed to be recoded to represent their predictions of repetition. For predictions about sequences ending in Red/Up outcomes, we took the raw response (between 0% and 100%) for participants in the continuous ("A") versions of each Study, and we coded Red/Up as "1" and Blue/Down as "0" for participants in the binary choice ("B") versions of each Study. For predictions about sequences ending in Blue/Down outcomes, we subtracted the raw response (between 0% and 100%) from 100 for participants in the continuous ("A") versions of each Study, and we coded Red/Up as "0" and Blue/Down as "1" for participants in the binary choice ("B") versions of each Study.

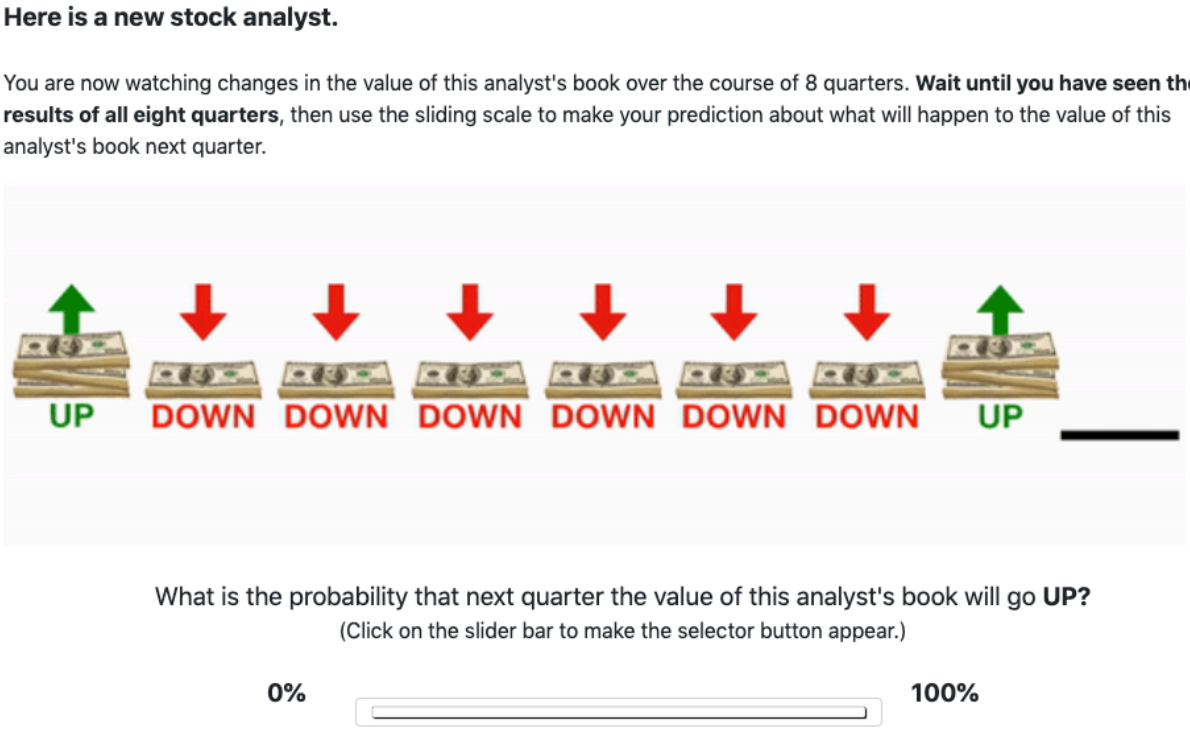
We found no significant differences between participants' predictions for sequences ending in Up or Down outcomes in the Analyst or Stock Conditions of any of our studies. We found no significant differences between participants' predictions for sequences ending in Red or Blue outcomes in the Bingo Conditions of Studies 1A, 1B, 2A, 3A, or 3B. We found a small difference between participants' predictions for sequences ending in Red or Blue outcomes in the Bingo50 Condition of Study 2B. The proportion of participants predicting repetition of Red streaks was slightly higher than 50% across all streak lengths, and the proportion predicting repetition of Blue streaks was slightly lower than 50% across all streak lengths.

Screenshots of Experimental Interface

Below are screenshots of the experimental interface used in each of the present studies. We present a sample of one Condition from each of the continuous ("A") versions of each Study. The experimental interface for the binary choice ("B") versions of each Study was identical, except for the question prompts and response inputs (refer to Figure B2, above).

Figure B3

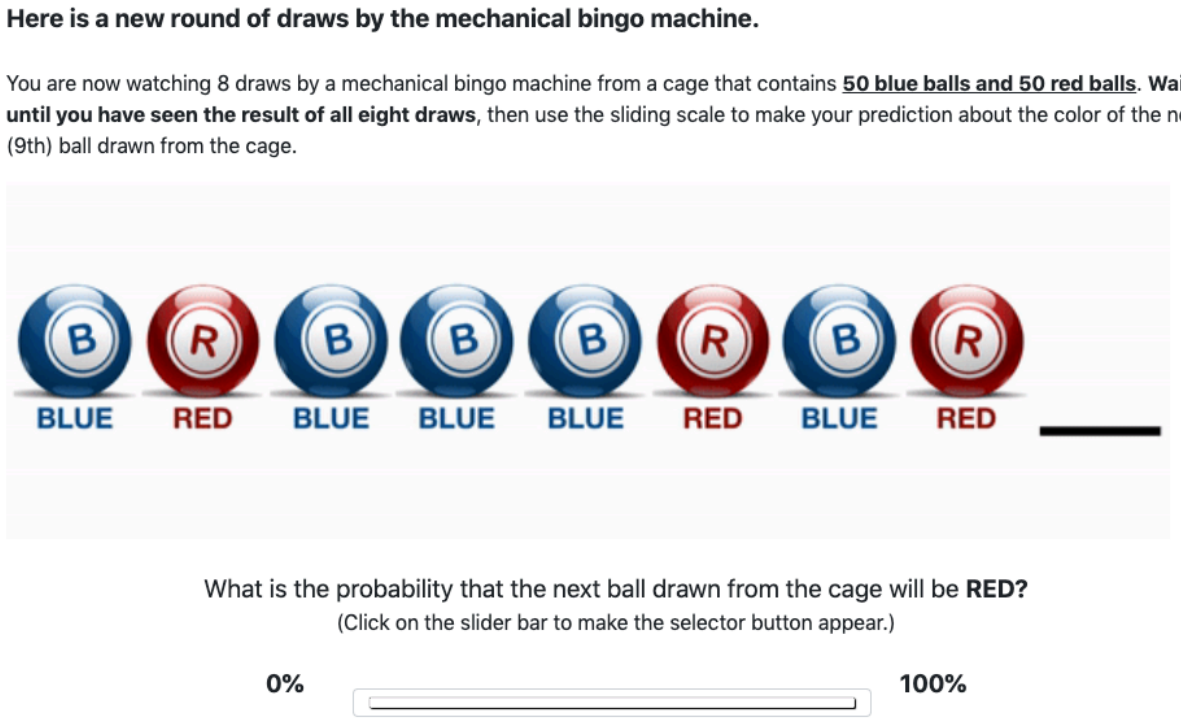
Screenshot of Experimental Interface for The AnalystUnknown Condition of Study 1A



Note: Images representing each outcome were revealed one at a time, from left to right, with a 1-second interval between the appearance of each image. In the Analyst Conditions of each of the present studies, Up outcomes were represented by an image of three stacks of dollar bills, with a green arrow pointing up above the stack, and the word "UP" below the stack. Down outcomes were represented by an image of one stack of dollar bills, with a red arrow pointing down above the stack, and the word "DOWN" below the stack. After all 8 stacks were revealed, a black bar beneath the 9th position in the sequence flashed three times. The selector button was hidden so that participants had to click on the scale to make it appear. This step was taken to avoid anchoring participants at any given point on the scale.

Figure B4

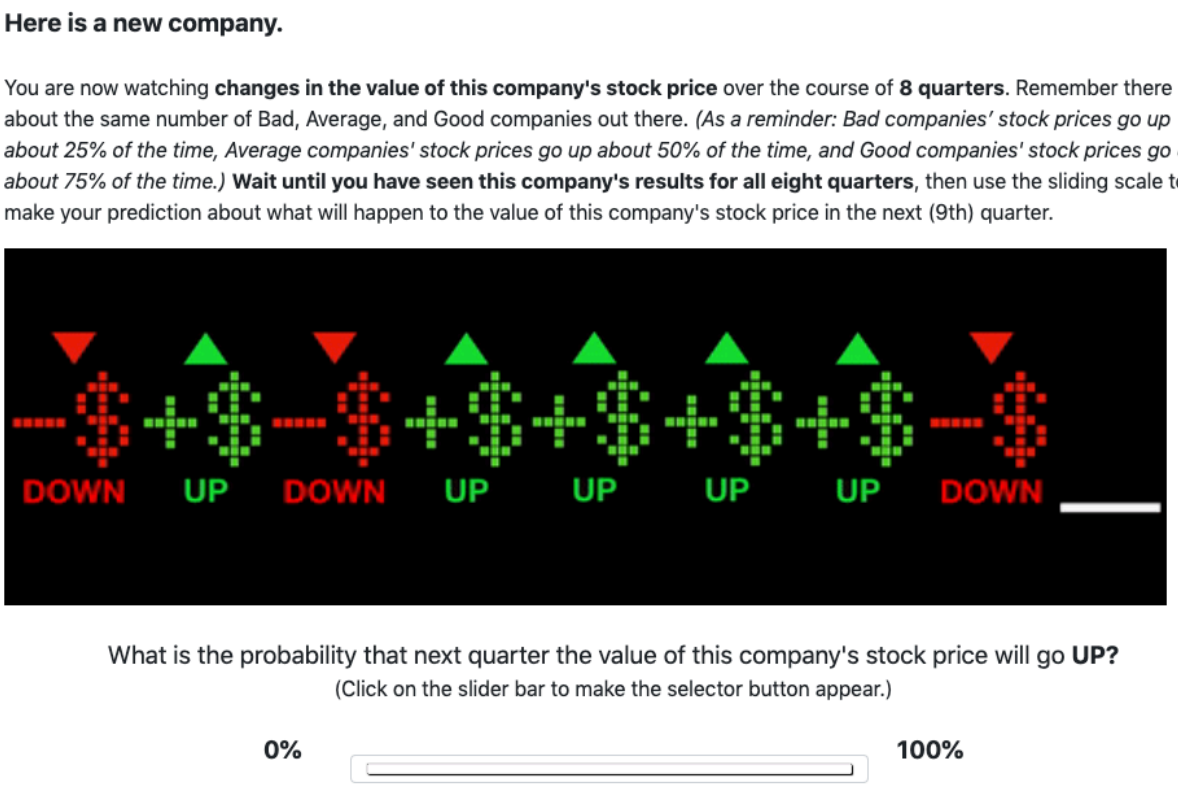
Screenshot of Experimental Interface for The Bingo50 Condition of Study 2A



Note: Images representing each outcome were revealed one at a time, from left to right, with a 1-second interval between the appearance of each image. In the Bingo Conditions of each of the present studies, Red outcomes were represented by an image of a red bingo ball marked with an "R" in its center, and the word "RED" below the ball. Blue outcomes were represented by an image of a blue bingo ball marked with a "B" in its center, and the word "BLUE" below the ball. After all 8 balls were revealed, a black bar beneath the 9th position in the sequence flashed three times. The selector button was hidden so that participants had to click on the scale to make it appear. This step was taken to avoid anchoring participants at any given point on the scale.

Figure B5

Screenshot of Experimental Interface for The Stock25-50-75 Condition of Study 3A



Note: Images representing each outcome were revealed one at a time, from left to right, with a 1-second interval between the appearance of each image. In the Stock Conditions of each of the present studies, Up outcomes were represented by an image of a green plus sign next to a green dollar symbol, with a green arrow pointing up above the dollar symbol, and the word "UP" below the dollar symbol. Down outcomes were represented by an image of a minus sign next to a red dollar symbol, with a red arrow pointing down above the dollar symbol, and the word "DOWN" below the dollar symbol. After all 8 symbols were revealed, a white bar beneath the 9th position in the sequence flashed three times. The selector button was hidden so that participants had to click on the scale to make it appear. This step was taken to avoid anchoring participants at any given point on the scale.

Appendix C: Verbal Reports

In this section, we look at differences between participants' qualitative descriptions of their prediction strategies, which were collected at the end of the experimental procedure in each of the present studies ("What was your strategy for predicting what would happen next? What information did you use to make your prediction?"). Two independent raters (Rater #1 and Rater #2) classified each participant's qualitative description by assigning a single category according to the instructions (Table C2 at the end of this section). Raters were blind to our hypotheses, and to the Study in which each participant participated. Two questions were raised by the raters after they began their task. These questions, and the responses provided to raters, are presented in Table C3 at the end of this section.

Participants' responses were classified into one of 8 categories: 1) balancing outcomes; 2) guessing; 3) estimating a proportion or counting outcomes, including references to updating estimates; 4) momentum or increasing probability of one outcome over the other; 5) "following instructions" (often reported to justify an answer of 50% in Studies 2A and 2B); 6) deciding which "type" of generator produced the sequence, particularly with reference to the distribution of rates provided in Studies 3A and 3B (e.g. high- versus low-performing analysts); 7) performing some sort of weighting calculation that takes into account the different "types" of generators, particularly in Studies 3A and 3B; 8) "other" unclassifiable responses.

Raters #1 and #2 only agreed on 58% of their ratings (781 out of 1,351). We measured inter-rater reliability using Cohen's kappa, and obtained a kappa value of 0.40, which indicates "fair" agreement. A third independent rater (Rater #3) was asked to resolve the disagreements between Raters #1 and #2. Rater #3 was given only the 571 responses to which Raters #1 and #2 assigned conflicting categories, and was provided the same instructions and FAQs. Rater #3 was

told to read each participant's response, and to decide whether Rater #1 or Rater #2's category was a better fit, given the instructions. Therefore, Rater #3 was *not* independently assigning categories to each response; rather, Rater #3 was restricted to the two categories previously assigned by Raters #1 and #2, and selected which of these two categories was a better match to the response. Rater #3 did not know the identity of Raters #1 and #2, was blind to our hypotheses, and to the Study in which each participant participated.

Table C1 presents a heatmap of the combined ratings that include Rater #3's resolutions. Darker green cells indicate that a category was assigned more often within a given Condition (Analyst, Bingo, Stock), Response Type (Continuous, Binary), and Rate (Unknown, Stationary .50, Specified Distribution .25–.50–.75). There was considerable variety in these reports in every experimental condition, but we see references to updating proportions comprise more than 50% of the responses for all experimental conditions (categories labeled "Proportion" and "Momentum" in Table C1). This fits our conclusion that base rate (proportion) updating is the primary inference process underlying the pervasive positive recency prediction patterns in our experiments.

We also see that "balancing" reports are scattered across experimental conditions and occur at highest rates for the random device Bingo Cage sequences (30% and 18% for the fixed 50% base rate). This is consistent with our conclusion that Law of Small Numbers ("balancing out") reasoning is likely to occur for random devices. Unsurprisingly, self-reports of reasoning about "types" of generators (coding categories labeled "Type" and "Weighting") are common when "types" (low, medium, high generators) are mentioned explicitly in the experimental instructions, as in the 25-50-75 Base Rate Studies (3A and 3B).

Table C1

Summary of Verbal Reports

	Unknown Rate		Stationary 50 Rate		25-50-75 Rate		
	Binary (Study 1B)	Continuous (Study 1A)	Binary (Study 2B)	Continuous (Study 2A)	Binary (Study 3B)	Continuous (Study 3A)	Total
Analyst							
Total # Responses	95	50	97	52	98	50	442
Instructions	2%	4%	1%	15%	3%	6%	19
Guessing	17%	14%	31%	21%	13%	18%	86
Balancing	2%	8%	15%	6%	3%	0%	27
Type	2%	2%	2%	0%	31%	14%	42
Weighting	0%	2%	1%	2%	4%	8%	11
Proportion	19%	30%	21%	25%	16%	28%	96
Momentum	49%	30%	24%	29%	24%	20%	134
Other	8%	10%	5%	2%	5%	6%	27
Bingo							
Total # Responses	108	50	101	56	109	50	474
Instructions	0%	0%	0%	9%	0%	0%	5
Guessing	24%	26%	41%	27%	13%	18%	118
Balancing	13%	22%	30%	18%	6%	4%	74
Type	0%	0%	0%	0%	26%	34%	45
Weighting	0%	0%	0%	0%	0%	0%	0
Proportion	48%	36%	15%	25%	45%	26%	161
Momentum	10%	12%	10%	7%	6%	8%	42
Other	5%	4%	5%	14%	4%	10%	29
Stock							
Total # Responses	97	44	103	48	93	50	435
Instructions	0%	0%	0%	6%	3%	6%	9
Guessing	18%	18%	18%	25%	10%	16%	73
Balancing	4%	5%	19%	6%	1%	0%	30
Type	0%	0%	0%	0%	37%	20%	44
Weighting	0%	2%	0%	4%	0%	0%	3
Proportion	11%	14%	19%	15%	19%	46%	85
Momentum	60%	55%	39%	33%	27%	10%	168
Other	7%	7%	4%	10%	3%	2%	23
Total	300	144	301	156	300	150	1351

Note: Summary of verbal reports, Studies 1–3. Proportion of participant strategies assigned to each category, by Condition, Response Type, and Rate. Each count represented as a proportion

of the total number of responses within a given Condition, Response Type, and Rate. Darker green cells indicate the more popular categories within a given Condition–Response Type–Rate section. Lighter green cells indicate less popular categories.

Table C2

Instructions Provided to Raters

Category	Description
Balancing	Participant mentioned something related to balancing out the number of outcome A (versus B) outcomes. For example, they chose A-type (B-type) to compensate for too few A-type (B-type) outcomes, or they chose B-type (A-type) to compensate for too many A-type (B-type) outcomes. General comments about having "too many" or "too few" of A-type (B-type) outcomes. Any reference to the number of Red or Blue balls remaining, e.g. "they took out 7 red balls, so there were only 43 red balls left."
Guessing	Participant mentioned something about their "gut" reaction, their emotions, feeling or sensing that an outcome would occur, or simply guessing.
Proportion	Participant mentioned something about estimating the proportion of A-type (B-type) outcomes, or the ratio of A-type to B-type outcomes. Comments related to "counting the number of [Red, Up] or [Blue, Down] outcomes," and choosing based on the number of each outcome. Any comments about basing their prediction on the relationship between A-type and B-type outcomes.
Momentum	Participant mentioned something about a trend, or change in the proportion of A-type to B-type outcomes over time. For example, participant says something like, "the company seems like it is on a roll, so the share price will probably continue to increase," or "the analyst has been doing really well, so his book will probably increase in value." Any comments related to momentum, increasing skill, increasing performance, learning, improving. Any comments related to deceleration, decreasing skill or effort, decreasing performance, something bad happening that changed the probability of a good outcome.
Instructions	Participant quotes the task instructions. For example, "the instructions said the rate was always 50%, so I always chose 50%."
Weighting	Participant mentions some sort of weighted calculation. For example, "there were three types of companies, so I thought about the three success rates," "only a good or average analyst could have so many successes, so I picked a rate halfway in between the good and average rates," "if it looked like it could have been from the cage with 25 or 50 red balls, I guessed what might come next from either of those cages."
Type	Participant talks about the "type" of analyst, company, stock, investor, bingo cage. For example, "I tried to figure out which type of analyst it was, and predicted based on the most likely type," or "If it looked like it was a good company, I guessed what would happen next for a good company, if it looked bad, I guessed what would happen next for a bad company," or "I tried to figure out which cage the draws were from, and guess what would come next from that cage."
Other	Anything that cannot be classified into one of the above categories.

Table C3

FAQs Provided to Raters

Question: There are a lot of responses that vaguely talk about 'I looked at the pattern' or 'I predicted the probability of the next one' with no real specifics that fall into any one category. Would responses like these fall into that 'other' category?

> **Answer:** For things like "I looked at the pattern" I would generally go with either the "Proportion" or the "Momentum" buckets. If they mention anything about what happened "at the end" or "changes" then it should go in the "Momentum" bucket. Otherwise, you could randomize your categorization between "Proportion" and "Momentum" and that will still probably give us a good enough sense of what's going on.
>
> For things like "I predicted the probability of the next one" I would probably go with the "Guessing" bucket. But, whenever either of the above suggestions really feel too "forced" to you, feel free to use the "Other" bucket.

Question: If there is a combination of types, would that fall under the 'other' category? For example, someone might respond saying they 'tried to follow the trend and made a prediction based on probability when they could, but occasionally went with their gut.'

> **Answer:** If the participant mentions two strategies with no preference "Sometimes I did x, other times I did y," record the first one they mention. If they indicate a preference "I tried to do x as much as I could, but when I could not do x I did y," record the one they said they were "trying" to do.

TIMING: Try to limit yourself to less than 30 seconds for each response.