

# On Making Forecasts From Binary Sequences: Uncovering Implicit Cues

Jason W. Beckstead and Mark V. Pezzo  
University of South Florida

The purpose of this article is to examine the statistical characteristics of binary sequences with the aim of uncovering the implicit cues that people use when making forecasts of what comes next. Information theory was used to quantify these statistical characteristics. In 2 experiments people were presented with 100 intact sequences of 20 Xs and Os and simply asked to forecast whether the 21st event in each sequence will be an X or an O. Multilevel logistic regression models were used to estimate the odds associated with these forecasts under different experimental manipulations. In a third experiment people judged the forecastability of sequences in a paired-comparison task. The results from the first 2 experiments showed that third-order redundancy (i.e., information provided by knowledge of the preceding pairs of events) was the most salient cue influencing forecasts. Experiment 3 showed that judgments of forecastability were based on this cue as well. When examining intact sequences with the goal of forecasting what comes next, people are more sensitive to higher-order transitional probabilities than has been previously suggested.

**Keywords:** alternation rate, entropy, forecasting, redundancy, transitional probability

**Supplemental materials:** <http://dx.doi.org/10.1037/xge0000834.supp>

*I love humans. Always seeing patterns in things that aren't there.*

—The Doctor

In everyday life, people make predictions, or forecasts, from sequences of past events. Is a basketball player likely to make or miss the next free throw? Will the next spin of the roulette wheel land on red or black? Will the stock market go up or down tomorrow? Will one's next online date turn out to be a dream or a dud? Such forecasts can influence behavior. This raises an interesting question, what statistical characteristics of sequences do people use as cues when forecasting what the next event will be? For sequences with only two alternatives (binary sequences), two characteristics that come to mind are the base rates of the two alternatives and the alternation rate of the sequence (described below). But, are there other, perhaps more influential, statistical characteristics of such sequences that serve as implicit cues?

In this article we explore this question with two experiments wherein people were presented with 100 sequences of 20 Xs and Os like this one, XXXOOOXXXXXOXOOOXXOO and simply asked to forecast whether the 21st event in each sequence will be an X or an O. We employ multilevel logistic regression models to estimate the odds associated with these forecasts under different


experimental conditions. In a third experiment we assess the perceived forecastability of such sequences.

Our query was motivated by three lines of psychological research involving sequences. The first is work on perceived randomness that has examined alternation rate extensively. The second is early work in probability learning that hinted at the relevance of higher-order transitional probabilities (described below) when making forecasts. The third, and the most influential, is the application of information theory to sequences of events to quantify various aspects of their structure.

This article is not about perceived randomness, or why people show the gambler's fallacy (although we do refer to this literature as it pertains to alternation rate). Nor is it about mental models, or the evolutionary advantages of perceiving streaks in the environment. What it is about is the complex structure of seemingly very simple stimuli and the sensitivity that people can show to various aspects of this structure. The experimental paradigm we use is new, therefore direct comparisons of our findings to published results are not possible; there are no previous experiments that used our approach. This said, we do attempt to draw some connections between our findings and the lines of research that inspired our investigation.

Our approach to quantifying the structure of sequences is rooted in concepts from information theory (Shannon, 1948a, 1948b). In the 1950s and 1960s information theory was viewed by many psychologists as a panacea in their quest for modeling the human capacity to process information; it was a shiny new tool for linking stimuli with responses in a way that had not been done before. As Luce (2003) points out, things didn't play out as expected because Shannon's approach treats all elements within a class of stimuli as interchangeable, but in most psychological experiments the stimuli are not perceived independently of one another leading to sequential effects, such as contrast and assimilation, in the responses. In

This article was published Online First August 27, 2020.

 Jason W. Beckstead, College of Public Health, University of South Florida; Mark V. Pezzo, Department of Psychology, University of South Florida.

The data and analyses appearing in this article are original and have not been presented previously.

Correspondence concerning this article should be addressed to Jason W. Beckstead, College of Public Health, University of South Florida, 13201 Bruce B. Downs Boulevard, Tampa, FL 33612. E-mail: [jbeckste@usf.edu](mailto:jbeckste@usf.edu)

essence, the tool didn't fit the job and it was set aside by most psychologists.

Our use of information theory is quite different. We find some of its basic concepts useful for quantifying aspects of our stimulus materials so as to permit more informed research questions and interpretations of our data. Brunswik argued that detailed analysis of the stimuli presented in psychological experiments is as important as the analysis of the responses obtained (Brunswik, 1956). When viewed from a Brunswikian perspective, we are applying information theory to the ecological side of the lens but not to the subjective side.

The remainder of the article is laid out as follows: In the next section we provide some key terms and definitions for describing the statistical characteristics of sequences. We then briefly review some relevant research in the areas of perceived randomness and probability learning that inspired us to design and conduct the first two experiments. Next, we discuss in detail some aspects of information theory that guided the formation of our hypotheses and research questions. We then present three experiments and discuss their results. Finally, we offer some general discussion of the findings and their implications.

### Some Terms and Definitions

Consider this sequence of Xs and Os: XXXOOOXXXXX-OXOOOXXOO. The sequence is made up of 20 binary consecutive events; the events are indexed by  $t$ , their position in the sequence;  $t$  ranges from 1 to 20. Here an event type can be either an "X" or an "O." The number of Xs equals the number of Os (10 of each), so the base rate (BR) of this sequence is balanced (BR = .5). The sequence also contains runs, or streaks, of consecutive events in which the event type does not change. The shortest run length possible is one. In this sequence the number of runs is 8: XXX, OOO, XXXXX, O, X, OOOO X, OO. The number of times that a sequence alternates is the number of runs minus one. The sequence can be described by its alternation rate (AR) which is the number of alternations relative to the number of alternations possible, or simply, the number of runs minus one, divided by the number of events minus one. This sequence has an AR of  $(\text{runs} - 1)/(\text{events} - 1) = (8 - 1)/(20 - 1) = .368$ .

A truly random sequence is said to have an AR of .5 because the next event has an equal chance of being the same as, or the opposite of, the previous event. Sequences with ARs less than .5 are referred to as "streaky" because they contain more streaks than does a random sequence, while sequences with ARs greater than .5 are referred to as "flippy" because they alternate, or flip back and forth, more so than does a random sequence. As an aside, for all possible sequences of 20 events, the ARs closest to .5 are  $.474 = (9/19)$  and  $.526 = (10/19)$ .

A sequence can also be described in terms of its sequential conditional probabilities or *transitional probabilities* (TPs). This type of probability specifies the relative frequency that X follows X, or that O follows X, for example. To illustrate precisely what we mean, the example sequence is examined two consecutive events at a time, that is, using an *observation window* of size two (i.e., events  $t$  and  $t - 1$ ) and there are 19 observations to be made. The observation window starts with events 1 and 2 and then moves to events 2 and 3, then to 3 and 4, and so forth, through events 19 and 20, tallying up the number of times that X, or O, follows X.

Because the observation window is size two, we refer to these as *second-order transitional probabilities*, or TP2s. For example, in this sequence there are 6 instances where X followed X, and 4 instances where O followed X, so the TP2s are:  $p(X \text{ on event } t | X \text{ on event } t - 1) = 6/10 = .600$ , and  $p(O \text{ on event } t | X \text{ on event } t - 1) = 4/10 = .400$ . Similarly,  $p(X \text{ on event } t | O \text{ on event } t - 1) = 3/9 = .333$  and  $p(O \text{ on event } t | O \text{ on event } t - 1) = 6/9 = .667$ . These calculations are illustrated in the upper portion of Table 1.

TP2s are related to AR in the following manner. AR is a weighted average of the two TP2 values that express the probability of change from one consecutive event to the next (i.e., from event  $t - 1$  to event  $t$ ). Each TP2 is weighted by the relative frequency of its previous event (i.e., how often it appears when the observation window size two). So,  $AR = [p(O \text{ on event } t - 1)][p(X \text{ on event } t | O \text{ on event } t - 1)] + [p(X \text{ on event } t - 1)][p(O \text{ on event } t | X \text{ on event } t - 1)]$ . Thus, in the example sequence,  $AR = (.474)(.333) + (.526)(.400) = .368$ .

The concept of a TP is easily extended to higher-orders by expanding the observational window. Setting the observational window to three events ( $t - 2$ ,  $t - 1$  and  $t$ ), the *third-order transitional probabilities*, TP3s, referring to probabilities that X or O follows a particular pair of consecutive events, can be obtained. There are four possible pairs that X and O can follow (XX, XO, OX, and OO); however, the number of observations decreases from 19 to 18 due to the increased size of the observation window. Likewise expanding the observational window to four events yields *fourth-order transitional probabilities*, TP4s. There are eight possible triples that X and O can follow but the number of observations drops to 17 (see the middle portions of Table 1 for illustrative calculations).

Under the assumption of ergodicity (i.e., that the probabilistic relations that characterize a sequence are constant), the TPs can be used to estimate the probability that the 21st event in the sequence will be an X or an O. We focus on estimating the probability of the 21st event being the opposite of the 20th to align these probability estimates with the interpretation of AR. The example sequence ends with O so we select the TP2 value for X following O and have  $p(21\text{st will be X} | \text{final observation is O}) = .333$ . The sequence may also be described as ending with the pair, OO, so we select the TP3 value for X following OO and have  $p(21\text{st will be X} | \text{final observation is OO}) = .400$ . Likewise the sequence may be considered as ending with the triple, XOO, and we select this TP4 value and have  $p(21\text{st will be X} | \text{final observation is XOO}) = .000$ .

For each of the 100 sequences to be presented in Experiments 1 and 2, we performed the calculations illustrated in Table 1. The AR and the three TP estimates that the 21st event will be the opposite of the 20th will be treated as independent variables in regression models of subjects' forecasts.

### Speculation

We are not suggesting that people are consciously performing the mathematical calculations of TPs when inspecting a sequence but rather that they are sensitive to and able to store such quantitative aspects of a sequence in (working) memory for use when making forecasts. It may be that when scanning a sequence (from left to right) to look for patterns, people encode not only which

Table 1

Example Calculation of Alternation Rate and Transitional Probabilities in a 20-Event Binary Sequence

Sequence ID: d1	XXXXXXXXXXOXOOOXOO				No. of runs = 8	
Alternation rate = (runs - 1) / (events - 1) = (8 - 1)/(20 - 1) = .368						
Observation window is 2 events						
Previous event ( $t - 1$ )	Event ( $t$ )		totals	$P$ (previous)	Transitional probabilities (TP2s)	
	X	O			$p$ (X   previous)	$p$ (O   previous)
X	6	4	10	.526	.600	.400
O	3	6	9	.474	.333	.667
	Total observations		19	1.000		
Observation window is 3 events						
Previous pair ( $t - 2, t - 1$ )	Event ( $t$ )		totals	$P$ (previous)	Transitional probabilities (TP3s)	
	X	O			$p$ (X   previous)	$p$ (O   previous)
XX	4	2	6	.333	.667	.333
XO	1	3	4	.222	.250	.750
OX	1	2	3	.167	.333	.667
OO	2	3	5	.278	.400	.600
	Total observations		18	1.000		
Observation window is 4 events						
Previous triple ( $t - 3, t - 2, t - 1$ )	Event ( $t$ )		totals	$P$ (previous)	Transitional probabilities (TP4s)	
	X	O			$p$ (X   previous)	$p$ (O   previous)
XXX	2	2	4	.235	.500	.500
XOX	0	1	1	.059	.000	1.000
XXO	1	1	2	.118	.500	.500
XOO	0	2	2	.118	.000	1.000
OXX	1	0	1	.059	1.000	.000
OOX	1	1	2	.118	.500	.500
OXO	0	2	2	.118	.000	1.000
OOO	2	1	3	.176	.667	.333
	Total observations		17	1.000		
Estimating the probability of the 21st event in the sequence under ergodicity						
TP2	$p$ (21st will be X   final observation is O) = .333					
TP3	$p$ (21st will be X   final observation is OO) = .400					
TP4	$p$ (21st will be X   final observation is XOO) = .000					
Weighted estimate	$P$ (21st will be X) = (.333 $\times$ 9 + .400 $\times$ 5 + .000 $\times$ 2)/(9 + 5 + 2) = .313					

Note.  $P( )$  represents a simple probability,  $p( )$  represents a conditional probability.

event type follows another, but also which event type follows different pairs or different triples. If people are encoding pairs and what follows them, then between-sequence differences in the TP3 estimates of what comes next should predict the forecasts that they make. Similarly, if they are encoding triples and what follows them, then TP4 estimates should predict their forecasts.

Just how the TP2, TP3, and TP4 estimates might be combined mentally is unknown. If people use them at all when making forecasts, perhaps they just average them (see Dawes, 1979). But, these estimates are not based on the same amount of data and so they are not equally reliable. One objective way to combine the three estimates is to use a weighted average, where each TP estimate is weighted by the number of times that its final observation (i.e., previous single, pair, or triple) appeared in the sequence (see the bottom portion of Table 1 for example calculation). This approach made sense to us from an encoding point of view, and its predictive utility will be assessed in Experiment 1.

## Research on Alternation Rate

AR has been studied extensively in the context of perceived randomness. Typically, sets of intact sequences are presented to subjects who are then asked to order sequences according to their apparent degree of randomness (Falk & Konold, 1997), to select the most random (Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993; Wagenaar, 1970; Wiegersma, 1987), or the least random (Green, 1982; Konold et al., 1993), to rate the randomness of each sequence (Scholl & Greifeneder, 2011) or to classify a given sequence as random or not (Altmann & Burns, 2005; Scholl & Greifeneder, 2011).

People tend to identify randomness with an excess of alternations (see Bar-Hillel & Wagenaar, 1991; Oskarsson, Van Boven, McClelland, & Hastie, 2009 for reviews). The AR most often perceived as random lies between .60 and .70 rather than at .50 (Falk, 1981; Falk & Konold, 1997; McDonald, 2009; Scholl & Greifeneder, 2011) but has been found to be as high as .81

(Alberoni, 1962) or higher, for example, when visual aesthetics rather than strict randomness is being judged (Sanderson, 2012).

From our point of view, the most relevant finding from this area of research is that when viewing intact sequences, people are sensitive to differences in AR. What has not been addressed, however, is the extent to which people use AR as a cue when making discrete forecasts of what comes next. We address this in our experiments.

### Transitional Probabilities and Classic Probability Learning Studies

Probability learning experiments were a prominent research paradigm in the 1950s and 1960s. In the typical experiment the subject is presented with a very long sequence of binary alternatives, presented one at a time. On each trial the subject is asked to predict what the next event type will be and then receives feedback. The focus in such studies is often on how quickly (slowly) subjects can learn the statistical nature of such sequences under various experimental conditions and to alter his or her prediction strategies to maximize accuracy.

Hake and Hyman (1953) were the first to examine the role of TPs in experiments on probability learning. Subjects were not only able to adjust their predictions to match the base rates of event types in sequences but also to adjust their prediction strategies to the TP2 values of sequences. Most interesting (to us) was the discovery that subjects were also responding to the TP3s as well. That is, trial-by-trial predictions were based in part on preceding pairs of stimulus values. In another early study, Anderson (1960) showed that trial-by-trial predictions were also related to preceding triples (TP4s) in the stimulus sequence. Research in probability learning continues; however, it was these early articles that led us to speculate that people might be sensitive to higher-order TPs when making forecasts from *intact* sequences rather than when making trial-by-trial predictions involving feedback.

The key distinction, as we see it, between our approach and the probability learning paradigm is that when presented with intact sequences, subjects do not have to “spend trials” to learn relative frequencies or conditional relations; all statistical characteristics of a sequence are laid out for inspection. Further, the statistical characteristics of an intact sequence are not entangled with any sequential dependencies among responses which are inherent in trial-by-trial forecasting.

### Applying Information Theory to Binary Sequences

How much information in a sequence is available for use when making a forecast?

Concepts from Shannon’s information theory (Shannon, 1948a, 1948b) can be used to answer this question. Shannon’s ideas were introduced to psychology by Miller and Frick (1949), and their applications to the field were further developed by Attneave (1959) and Garner (1962). Shannon used the term *entropy*, but psychologists have preferred the term *uncertainty*. Uncertainty is a single number associated with a finite probability distribution. Uncertainty is typically quantified in binary digits, or bits. When uncertainty is reduced, it becomes information. Or, one might say that information reduces uncertainty. So, once we can quantify uncertainty, we can also quantify information in a similar manner.

When applying information theory to sequences we find Attneave’s (1959) treatment quite instructive, and so we borrow some terminology. The term *redundancy* refers to the amount of information provided with regard to forecasting what the next symbol or event type will be. All sequences may be characterized by some degree of redundancy (from 0% to 100%). No redundancy (0%) means that all the symbols have an equal probability of occurrence, and nothing that we may know about the history of the sequence makes the next symbol any more predictable. At the opposite extreme (100% redundancy), the symbols appear in a perfectly lawful and regular pattern, making it possible to predict with complete certainty what the next symbol will be. A sequence may be described by different orders of redundancy. *First-order redundancy* depends solely on the probabilities of the individual symbols appearing in the sequence (i.e., the BR). Higher orders of redundancy refer to sequential dependence within a sequence. *Second-order redundancy* is the extent to which the next symbol is predictable by knowledge of the immediately preceding symbol. How predictable the next symbol is by knowledge of the immediately preceding *pair* of symbols defines *third-order redundancy*. By extension, *fourth-order redundancy* refers to how predictable the next symbol is based on knowledge of the immediately preceding *triple* of symbols.

The uncertainty and redundancy values associated with a sequence may be calculated using the BR and the TP distributions. We illustrate the calculations using the example sequence presented in Table 1. For the BR, we have only two probabilities to consider:  $P(X) = 10/20 = .5$  and  $P(O) = 10/20 = .5$ . Each probability is multiplied by the  $\log_2$  of its reciprocal and these two products are summed:  $(.5)\log_2(1/.5) + (.5)\log_2(1/.5) = 1.0$  bits. This quantity is referred to as *first-order conditional uncertainty*,  $U_1$ . It is the amount of uncertainty remaining in our forecasting task after considering the base rate of the two event types. The *unconditional uncertainty* in the forecasting task,  $U_o$ , is simply the  $\log_2$  of the number of options (an X or an O in this case), so  $\log_2(2) = 1$  bit. By subtraction we obtain first-order redundancy, or the amount of information gained by knowledge of the BR. Because  $U_o - U_1 = 1 - 1 = 0$  bits, the BR of this sequence provides no information about what the next event in the sequence might be.

For the higher-order conditional uncertainties we work with the matrices of transitional probabilities. The calculation can be accomplished in three steps. We illustrate using the TP2 values as presented in Table 1. First, each of the four TP2 values is multiplied by the  $\log_2$  of its reciprocal. Second, within a row these products are summed and then weighted by  $P(\text{previous})$ . Third, the results from Step 2 are summed over rows. Working with the first row we have  $.526[(.600)\log_2(1/.600) + (.400)\log_2(1/.400)]$ , and for the second row we have  $.474[(.333)\log_2(1/.333) + (.667)\log_2(1/.667)]$ . These are summed to yield 0.946 bits. This quantity is *second-order conditional uncertainty*,  $U_2$ , which is the amount of uncertainty remaining in the forecasting task after considering previous single events. Again by subtraction, we obtain second-order redundancy, the amount of information gained by knowledge of the previous single events:  $U_1 - U_2 = 1 - 0.946 = 0.054$  bits.

The same three steps may be applied to the eight TP3s and to the 16 TP4s in Table 1 to determine how much uncertainty remains after considering previous pairs and triples in the example se-



quence, respectively. The amount of *third-order conditional uncertainty*,  $U_3$  (after considering previous pairs), is 0.909 bits. Finally, the amount of *fourth-order conditional uncertainty*,  $U_4$  (after considering previous triples), is 0.633 bits.

The estimated net amount of information in a sequence available for use in forecasting depends upon the highest order of uncertainty used when analyzing the sequence. In the present case we can refer to fourth-order information (redundancy), and use Attneave's symbol,  $C_4$ . This quantity may be expressed in two ways, first simply as  $C_4 = U_0 - U_4 = 1 - 0.633 = 0.367$  bits, and because  $U_0 = 1$  bit, we may say that redundancy amounts to about 36.7% at the fourth order. Alternatively,  $C_4 = (U_0 - U_1) + (U_1 - U_2) + (U_2 - U_3) + (U_3 - U_4)$  which gives the same result, 0.367 bits. In this form, the terms to the right of the equal sign indicate the relative contributions from the BR, previous single events, previous pairs, and previous triples, respectively. It is this definition that we find most useful because the net information available for forecasting can be broken down into additive parts. The amount of information gained by knowledge of the BR,  $I_{BR} = (U_0 - U_1) = 0$  bits. The average amounts of information gained by knowledge of the previous single event, previous pair, and previous triple in the example sequence are thus:  $I_{single} = (U_1 - U_2) = 0.054$  bits,  $I_{pair} = (U_2 - U_3) = 0.037$  bits, and,  $I_{triple} = (U_3 - U_4) = 0.277$  bits, respectively.

In the universe of 524,288 unique 20-event binary sequences there are 92,378 that have a balanced BR. The information analysis demonstrated above was conducted on each of them. Averaging across all these sequences, the amount of information from preceding single events is 0.041 bits ( $SD = 0.059$ ), from preceding pairs is 0.108 bits ( $SD = 0.100$ ), and from preceding triples is 0.228 bits ( $SD = 0.138$ ). For 87,764 (95%) the amount of information gained by knowledge of the preceding pairs or triples is greater than that gained by knowledge of the preceding single events. This implies that describing sequences simply in terms of their TP2s or their AR does not adequately distinguish among them.

### More Speculation

The information analysis above shows how the amount of information in a sequence may be decomposed into different sources. This raises the interesting possibility of the different sources favoring different forecasts. Consider the following scenario involving only first- and second-order redundancy. If a sequence has an unbalanced BR, then first-order redundancy will be greater than zero. For a 20-event sequence with 13 Xs and 7 Os, the BR may be expressed as .65 and  $I_{BR} = 0.07$  bits. Second-order redundancy may be calculated from AR. The amount of information associated with AR increases symmetrically moving away from .5. Therefore, if a sequence is either moderately streaky or moderately flippy (AR = .342, or AR = .658, respectively) then  $I_{single} = 0.07$  bits. If a sequence has both a BR = .65 and an AR = .342, each source contributes the same amount of information. Now for the interesting part; let's say that the sequence ends with an O. Because the sequence is streaky, the 21st event is most likely to be another O. Based on the BR, however, the 21st event is most likely to be an X (because X is the more common event type).

This line of thinking led us to an interesting question: How will people respond when presented with sequences in which the AR

and the BR each make available the same amount of information, but where these two sources of information point to different forecasts? More specifically, among moderately streaky sequences where the tendency is to forecast a repeat, will manipulating the BR to favor forecasting a flip negate or maybe even reverse this tendency?

Information theory also suggests another objective means of forming a weighted average of the three TP estimates that the 21st event in a sequence will be the opposite of the final observation. Each estimate is weighted by the relative amount of information provided by its source; the TP2 estimate is weighted by the amount of information provided by the single events, TP3 by that provided by pairs, and TP4 by that provided by triples. The predictive utility of this model will also be assessed in Experiment 1.

### Hypotheses and Research Questions

The experimental paradigm used in Experiments 1 and 2 borrows features from studies in perceived randomness and in probability learning. Several sequences (of 20 Xs and Os) were presented in their entirety in written format, but rather than ask subjects to judge the randomness of a sequence, they were asked instead to forecast what the next event type in each sequence is most likely to be. Unlike subjects in probability learning tasks, they did not receive any feedback because there is no "correct" answer for what the 21st event type is in any given sequence. Nonetheless, by systematically manipulating various statistical characteristics of several sequences we can determine which characteristics are associated with subjects' forecasts.

Because it has long been known that people have preferences for one symbol over another (see Goodfellow, 1938), we presented complementary pairs of sequences, formed by exchanging all Xs for Os, to control for such preferences. This meant that half of the sequences ended with X and half ended with O. Prior to analysis, responses were coded such that a 1 indicated that the subject was forecasting a flip, that is, that the 21st event would be the opposite of the 20th event type, and a 0 indicated that the subject was forecasting a repeat, that is, that the 21st event would be the same as the 20th event type. Complimentary pairs of sequences also allowed us to assess the reliability of each subject's forecasts.

Because our experimental approach is new and combines features from other lines of inquiry, we wanted to provide some basis for comparison; therefore, we begin by assessing the influences of AR and BR separately to establish that our approach yields intuitive results. We then assess the combined influence of AR and BR, and conclude our analysis by addressing the most speculative questions regarding the influence of higher-order TPs. Experiment 2 is essentially a replication of Experiment 1, but using a different set of sequences. Experiment 3 uses a qualitatively different approach intended to supplement the first two experiments and will be described in detail following the results of Experiment 2.

*Hypothesis 1:* Based on the findings from perceived randomness studies, moderately flippy sequences (ARs between .6 and .7) should be perceived as random and so the odds of forecasting a flip should be 1 to 1. Our hypothesis, based on symmetry in information availability, is different; the odds of forecasting a flip will increase as AR increases from .5 and will decrease as AR decreases from .5.

*Hypothesis 2:* Altering the BR of sequences, by nearly doubling the number of one event type relative to the other, provides information about what the next event type is likely to be. Therefore among AR-neutral sequences (where little to no information is provided by AR) an unbalanced BR will influence forecasts. Specifically, unbalanced BR sequences that end in the less common event type are expected to increase the odds of forecasting a flip, while unbalanced BR sequences that end with the more common event type are expected to increase the odds of forecasting a repeat.

*Question 1:* When presented with sequences in which the AR and the BR each make available the same amount of information, but where these two sources of information point to different forecasts, how will people respond? Specifically, among moderately streaky sequences where the tendency is to forecast a repeat, will manipulating the BR to favor forecasting a flip negate or maybe even reverse this tendency?

*Question 2:* Will the three types of TPs influence forecasts, and if so, what are their relative contributions to these forecasts?

## Experiment 1

### Method

**Statement of ethical approval.** This experiment and the two that follow were reviewed and approved by the university's Institutional Review Board and were conducted in accordance with the ethical standards of the American Psychological Association.

**Subjects.** One hundred fifty-two students, faculty, and staff on two campuses of a major university located in the southeastern United States volunteered to participate in this experiment.

**Materials.** One hundred binary sequences, each containing 20 events, were created using the symbols "X" and "O." Because some of our questions dealt with the influence that BR might have on forecasts, we wanted to use some sequences where BR was balanced (10 Xs and 10 Os) and so provided no clue as to what the next event type might be. This also allowed us to rule out the possibility of a probability matching strategy being applied to balanced BR sequences; with 21 events, for example, one outcome (an X or an O) would have to occur more frequently than the other. Further, we wished to discourage any aesthetic motive to achieve balance via forecasting the less frequently occurring event type. With 20 events (10 of each type) regardless of which type is

forecast, the sequence becomes unbalanced once the forecasted event is added to it.

There were three characteristics that we used to define each sequence: its BR (the number of Xs and Os), its AR, and its final event type (X or O). Six ARs ranging from moderately streaky to moderately flippy (.316, .368, .474, .526, .632 and .684) and two BRs, balanced (10 of each event type) and unbalanced (13 of one event type and seven of the other), were used.

According to Scholl and Greifeneder (2011), a potentially important characteristic of a sequence that can moderate the influence of AR on judgments is the maximum run length (MRL). Typically, MRL and alternation rate will be inversely related. We limited the MRL to five in the most streaky sequences (ARs of .316 and .368) and to three in the other sequences. Another potentially influential characteristic that we controlled for was the length of the final run in each sequence. Carlson and Shu (2007) provided evidence that a run length of three is the pivotal value in the perception of streaks. Therefore, none of the sequences ended in more than two of the same event type. This was done in an effort to distribute attention throughout the entirety of each sequence.

The experimental design contained 10 sequence types or conditions. Six conditions had a balanced BR, and four had an unbalanced BR (see Figure 1). With these design constraints in place, 50 sequences (five for each condition) were randomly selected from the universe of sequences. No attempt was made to match sequences on TP3s or TP4s. For each of these 50 unique sequences a complimentary version was fashioned by exchanging all Xs and Os. This yielded 100 unique sequences (10 sequences in each condition) for presentation. Within each condition five of the sequences now ended with an X and their five complements ended with an O.

Sequences were printed one per line, with spacing between lines set at 1.5 which permitted 25 sequences to appear on each of four consecutive pages. Within each sequence, the symbols were separated by a space. Each sequence was preceded by an ellipsis, "...," to convey the impression that the 20 events being presented were part of a longer sequence and to discourage people from thinking about how each sequence began.

**Counterbalancing.** Matthews (2013) found that when intact sequences were presented as a group and subjects were asked to rate the likelihood that the next event in each sequence would be a head or a tail, ratings of each sequence were influenced by the sequence that preceded it. In an effort to minimize such context effects, the 100 sequences in the current study were presented

		Alternation Rate					
		.316	.368	.474	.526	.632	.684
Balanced Base Rate (10/10)	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	
	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O
Unbalanced Base Rate (13/7)	Condition 7	Condition 8	Condition 9	Condition 10	Each subject responded to 100 sequences.		
	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O	10 sequences 5 end in X 5 end in O			

Figure 1. Within-subject design for Experiments 1 and 2.

using an extensive within-subject and between-subjects counterbalanced design.

To start we arranged the 100 sequences into five sets of 20 sequences. Each set contained one sequence from each of the 10 conditions and its complement. The sequence types within each set of 20 were ordered for presentation so as to appear to the subject as unsystematic as possible. From the subject's perspective, we felt that the most obvious experimental manipulation would likely be the event type (X or O) with which the sequence ended when sequences were presented one after another on a page, so this was addressed first. We thought about simply alternating such sequences but felt that this might focus attention on the last event at the expense of attending to each sequence in its entirety. We also considered ordering them so that sequences ending with X or O would appear in truly random order, that is with an AR of .50; however, in anticipation that subjects would perceive this ordering as streaky, the sequences within each set of 20 were ordered so that the AR of the last event type across sequences was slightly flippy (.579). Second and working within this constraint, the order of sequence types within each set of 20 was further manipulated to minimize the possible context effects arising from AR and BR values of the preceding sequence. The autocorrelation of the AR values across the sequences was  $-.237$  and the autocorrelation of the BR values across the sequences was  $.125$ . The net result of these within-subject counterbalancing steps was that the autocorrelation of the TP2 values (which incorporates both the AR and the BR of a sequence) across sequences was nearly zero (.006).

Because the task involved examining 100 sequences without feedback we were concerned about the possibility of warm-up and fatigue effects. To address these concerns the five sets of 20 were arranged in five orders such that each set fell in each of five positions within the total 100. That is, labeling the sets 1, 2, 3, 4, and 5, the orders: 1, 2, 3, 4, 5; 2, 3, 4, 5, 1; 3, 4, 5, 1, 2; 4, 5, 1, 2, 3; and 5, 1, 2, 3, 4 were used. From these orders, five versions of a questionnaire were printed and subjects were randomly assigned to complete one of the five. This between-subjects counterbalancing permitted assessment of warm-up and fatigue effects as well as the degree of interchangeability among the five sets of 20 sequences.

**Procedure.** Data were collected in classroom settings using a written questionnaire. Each individual received a large chocolate bar for participating.

**Task instructions.** The task was presented as one of detecting patterns in sequences. We indicated that the sequences were created by mathematical algorithms in an effort to instill a belief that there were in fact patterns to be detected and to encourage examination of each sequence in its entirety. Subjects were provided with the following description of the task:

We are interested in understanding how people are able to distinguish randomness from subtle patterns, or regularities, in sequences of events. The way that we do this is to present individuals with a part of a sequence and ask them to "fill in" the next symbol in the sequence. Some regularities are easier to notice than others.

On the next four pages you will be presented with 100 sequences of 20 symbols each. **Your task is to examine each sequence carefully and predict (to the best of your ability) which symbol you think is most likely to occur next.** We used various mathematical rules (algorithms) to construct these sequences. Some sequences were made

using such complex algorithms that it is nearly impossible to predict what the next symbol will be. Regardless, we want you to **try your best** in each case.

The sequences are presented in no particular order; they do not get any easier or any more difficult as you move through the pages. Please provide a response following each sequence, even if you have to make a kind of guess as to what you think the next symbol will be. For each sequence, please write an "X" or an "O" in the blank that follows to indicate which symbol you think is most likely to occur next.

**Analytic approach.** Because our experimental paradigm is new and our aim is one of discovery, we felt it important to maximize the signal-to-noise ratio in the data prior to conducting substantive analyses. The design of the experiment permitted the calculation of Cohen's kappa for each subject using their responses to the 50 pairs of complementary sequences. Kappa corrects for chance agreement due to differences in the marginal distributions of responses. This meant that two subjects who differed in the marginal distributions of their forecasts (e.g., subject 1 may have forecasted an equal number of flips and repeats, while subject 2 a preponderance of repeats) could be equated in terms of their consistency. Although the value of kappa and its standard error may be different for each subject, the resulting *t* statistic puts them all on the same metric for comparison (because the number of cases being judged is constant). As a means of maximizing the signal-to-noise ratio in the data, only subjects with kappa values significantly greater than chance were retained for primary analysis.

Substantive analyses were done using multilevel logistic regression models conducted with SPSS GENLINMIXED procedure specifying a binomial distribution and the logit link function. Prior to conducting analyses, responses to complementary sequences were recoded so that for all sequences a response of 1 indicated that the subject was forecasting a flip, that is, that the 21st event would be the opposite of the 20th event type, and a 0 indicated that the subject was forecasting a repeat, that is, that the 21st event would be the same as the 20th event type.

All models included a random intercept to account for the nested structure of the data (100 responses were nested within each subject) as well as fixed and random effects for each predictor variable. The random intercept term captures individual differences in the number of flip forecasts. Including a random effect for a predictor variable allows the influence of the predictor (i.e., its slope) to differ across individuals. The model provides an estimate of the variance in these slopes and a test of significance on its magnitude. Assuming that these slopes are normally distributed in the population, the square root of this variance can be used to estimate the proportion of the population that has a slope above, or below, any meaningful cutoff, such as 0.00. Continuous predictor variables (AR and TPs) were either mean-centered or standardized. The predictive utility (goodness of fit) of each logistic model was assessed in two ways: first, using area under the receiver operating characteristic curve (aROC), and second, using the correlation of the observed responses with the model-predicted responses.

## Results

**Data quality.** Data were obtained from 152 subjects. The average time to complete the task was 40 min. Responses from each subject were screened by checking the frequencies of Xs and

Os and response alternation rates against expected values based on the binomial distribution. Data from two subjects were discarded because these individuals apparently did not take the task seriously. One was excluded because he indiscriminately wrote alternating Xs and Os throughout the 100 sequences. The other was excluded because she always responded with the opposite of the last event, regardless of the AR and BR. This left data from 150 individuals for analysis. Most were students and the majority were female (66.7%). The mean age was 28.3 years,  $SD = 11.3$ .

To assess warm-up and fatigue effects we compared the frequencies of X responses and the alternation rate of responses over the five consecutive blocks of 20 sequences (aggregating across the five sets). The average number of X responses per block showed minimal variation (range 9.77 to 10.30). The average alternation rate in responses per block ranged from .547 to .569. Based on these analyses warm-up and fatigue effects were not present in our data. We also examined the frequencies of X responses and the alternation rate of responses for each set of 20 sequences (aggregating across the five orders) to address our assumption of interchangeability among the sets. The average number of times subjects responded X within a set ranged from 9.97 to 10.37, and the average alternation rate in responses across

sets ranged from .529 to .574. Based on these analyses we felt that it was appropriate to combine responses from the five sets for substantive analyses.

The final step in assessing the quality of our data involved calculating Cohen's kappa on the responses of each subject over the 50 complementary pairs of sequences. Using a criterion of  $p < .05$  to define greater than chance within-subject consistency, 85 subjects had kappa values significantly higher than chance and were retained for primary analysis to maximize the signal-to-noise ratio in the data. By definition, the remaining 65 contribute more noise than signal to the analyses but will be examined subsequently.

#### Substantive analysis.

**Hypothesis 1.** The odds of forecasting a flip will increase as AR increases from .5 through .684, and will decrease as AR decreases from .5 through .316. This was tested using 60 responses per subject under Conditions 1 through Condition 6 (5,100 forecasts in total). AR values were mean-centered (around .5) for analysis. The results are shown as Model 1 in Table 2,  $F(1, 5098) = 253.583, p < .001$ .

In Model 1 the intercept was  $-0.105$  and the slope for AR was  $4.318$  and was significant, providing support for Hypothesis 1. The

Table 2

*Fixed Effect (Top) and Random Effect (Bottom) Estimates for Models Predicting Forecast Decisions (Experiment 1)*

Model description	Fixed effects					Model fit statistics	
	Term	Coefficient	SE	t	p	aROC	$r_{\text{obs,prd}}$
1. Influence of AR when BR is balanced Conditions 1–6	Intercept	−0.105	0.068	−1.549	.122	.719	.383
	AR	4.318	0.271	15.924	.001	[.705, .732]	[.359, .406]
2. Influence of BR when AR is neutral Conditions 3, 4, 9, and 10	Intercept	−0.858	0.086	−9.976	.001	.804	.525
	End in MC	−0.319	0.116	−2.752	.006	[.789, .818]	[.500, .549]
	End in LC	1.764	0.147	12.021	.001		
	Last 2 Same	1.143	0.116	9.829	.001		
3. Influence of BR when in opposition to AR Conditions 1, 2, 7, and 8	Intercept	−1.819	0.117	−15.545	.001	.803	.528
	End in LC	1.482	0.102	14.514	.001	[.788, .817]	[.503, .552]
	Last 2 Same	1.739	0.125	13.966	.001		
4. Relative influence of SCPs Conditions 1–6 (predictors standardized)	Intercept	−0.110	0.077	−1.424	.154	.821	.558
	TP2	0.075	0.044	1.736	.083	[.810, .832]	[.539, .577]
	TP3	0.870	0.069	12.712	.001		
	TP4	0.281	0.055	5.139	.001		
	Random effects						
	Term	Variance	SE	z	p		
1. Influence of AR when BR is balanced Conditions 1–6	Intercept	0.314	0.064	4.928	.001		
	AR	1.645	0.961	1.712	.087		
2. Influence of BR when AR is neutral Conditions 3, 4, 9, and 10	Intercept	0.304	0.086	3.547	.001		
	End in MC	0.078	0.127	0.617	.537		
	End in LC	0.890	0.262	3.395	.001		
	Last 2 Same	0.317	0.125	2.538	.011		
3. Influence of BR when in opposition to AR Conditions 1, 2, 7, and 8	Intercept	0.413	0.108	3.832	.001		
	End in LC	0.086	0.058	1.494	.135		
	Last 2 Same	0.602	0.175	3.433	.001		
4. Relative influence of SCPs Conditions 1–6 (predictors standardized)	Intercept	0.409	0.081	5.029	.001		
	TP2	0.034	0.023	1.527	.127		
	TP3	0.175	0.052	3.371	.001		
	TP4	0.076	0.033	2.334	.020		

*Note.* AR = alternation rate; aROC = area under the receiver operating characteristic curve; BR = base rate; MC = more common event type; LC = less common event type; TP = transitional probability. 95% CIs are shown in brackets.



random effect for this slope was not significant (meaning that the size of the slope did not vary significantly across subjects). Logistic regression models provide estimates of the logit of a binary response variable. The logit may be transformed to provide estimates of the odds of responding one way or the other at different values of the predictor variable(s).

In this model if we exponentiate the logit we obtain the estimated odds of forecasting that the next event will be the opposite of the final event (i.e., the odds of forecasting a flip). To illustrate, consider truly random sequences with an  $AR = .5$ . As  $AR$  was mean centered, a value of  $.5$  becomes 0, and so the estimated logit is  $-0.105 + (.5 - .5)(4.318) = -0.105$ . Exponentiating this value, we obtain 0.90, meaning that the estimated odds of forecasting a flip are 0.90 to 1. As the intercept was not significantly different from zero, this means that the odds (0.90) were not significantly different from 1, so the model predicts that people are equally likely to forecast a flip or a repeat when faced with truly random sequences.

For comparison, flippy sequences with  $AR$  of  $.684$  have an estimated logit of  $-0.105 + (.684 - .5)(4.318) = 0.6895$  and exponentiating gives an odds of 1.99. To compare forecasts at two different values of  $AR$ , we form an odds ratio ( $OR$ ) =  $1.99/0.90 = 2.21$ . So, for moderately flippy sequences the odds of forecasting a flip are 2.21 times greater than they are for a truly random sequence.

For streaky sequences with  $AR$  of  $.316$  the estimated logit is  $-0.105 + (.316 - .5)(4.318) = -0.8995$ , and exponentiating gives an odds of 0.41. Comparing this estimate with a truly random sequence, the  $OR = 0.41/0.90 = 0.46$ . So, for moderately streaky sequences the odds of forecasting a flip are less than half of what they are for a truly random sequence. When  $OR$ s are less than 1, we find it helpful to work with their reciprocals. Thus, while the  $OR$  for forecasting a flip are 0.46 for such streaky sequences, the  $OR$  for forecasting that the next event will be the *same* as the final event (i.e., forecasting a repeat) are  $2.20 = (0.90/0.46)$ .

**Hypothesis 2.** Among  $AR$ -neutral sequences, an unbalanced  $BR$  will influence forecasts. Specifically, unbalanced  $BR$  sequences that end in the less common event type will increase the odds of forecasting a flip, while unbalanced  $BR$  sequences that end with the more common event type will increase the odds of forecasting a repeat.

Forty responses per subject under four experimental conditions (3, 4, 9, and 10) were involved in testing this hypothesis (3,400 forecasts in total). All sequences had neutral  $AR$ s (.474 or .526, which are as close to  $.5$  as is possible for sequences of 20 events). Half of the sequences had a balanced  $BR$  and half had an unbalanced  $BR$ . Two dummy variables were used to code our manipulation of  $BR$ . Unbalanced  $BR$  sequences that ended with the more common event type were coded 1 on the first dummy variable, unbalanced  $BR$  sequences that ended with the less common event type were coded 1 on the second dummy variable, and sequences with a balanced  $BR$  served as the reference group (coded 0, 0). This coding strategy provided two significance tests for our  $BR$  manipulation. For the hypothesis to be supported both tests must be significant and the coefficient for sequences ending with the less common event type must have a positive sign while the coefficient for the sequences ending with the more common event type must have a negative sign.

The coefficients for both dummy variables were significant, however both had positive signs; this gave us pause. To figure out why this occurred, we conducted a detailed inspection of the 100 sequences that revealed that the number of sequences ending with two events of the same type was not distributed equally across conditions. We therefore created a dummy variable to code for this characteristic and found that it was positively correlated with the response variable. Furthermore, this variable was negatively correlated with the dummy variable anticipated to have a positive coefficient, and positively correlated with the dummy variable anticipated to have a negative coefficient. Epidemiologists refer to such variables as confounders and recommend including them in the model to statistically correct for their influence(s). The model testing Hypothesis 2 and including the confounder is shown in Table 2 as Model 2,  $F(3, 3396) = 77.692, p < .001$ . After controlling for the confounder, the coefficient for unbalanced sequences ending with the less common event type was positive (and significant) and the coefficient for those ending with the more common event type was negative (and significant) providing support for the hypothesis. The random effect for sequences ending with the more common event type was not significant, but the random effect for sequences ending with the less common event type was significant, indicating that the slopes for this predictor varied across subjects. Assuming that these slopes are normally distributed in the population, 97% of the subjects were estimated to have slopes greater than zero.

The adjusted odds of forecasting a flip were 0.75 for the balanced  $BR$  sequences. For unbalanced  $BR$  sequences ending with the less common event type, the adjusted odds were 4.38, and for unbalanced  $BR$  sequences ending with the more common event type the adjusted odds were 0.55. Using the balanced  $BR$  sequences as a reference, subjects were over 5 times more likely to forecast a flip when the  $BR$  manipulation favored this forecast ( $OR = 4.38/0.75 = 5.84$ ). Subjects were 1.36 times more likely to forecast a repeat when the  $BR$  manipulation favored this forecast ( $OR = 0.75/0.55 = 1.36$ ). Taken together these results demonstrate that people are sensitive to changes in the  $BR$  of  $AR$ -neutral sequences and can use this cue (as predicted) when making their forecasts. What is not clear to us at this point is why this effect was not symmetric; the manipulation of  $BR$  had a much larger effect when it favored forecasting a flip versus when it favored forecasting a repeat.

**Question 1.** Among moderately streaky sequences where the tendency is to forecast a repeat, will manipulating the  $BR$  to favor forecasting a flip negate or maybe even reverse this tendency?

This question was addressed using 40 responses per subject in Conditions 1, 2, 7, and 8 (3,400 forecasts in total). Sequences in Conditions 1 and 7 had  $AR = .316$ , and those in Conditions 2 and 8 had  $AR = .368$ . The average  $AR$  for these sequences is thus  $.342$ , which meant that on average they made available 0.07 bits of information favoring a forecast of repeat. Sequences in Conditions 7 and 8 had an unbalanced  $BR$ , which meant that they made available an additional 0.07 bits of information; however, because they ended with the less common event type, this information favored a forecast of flip. We ran this analysis both with, and without, the confounder. In both cases the results led to the same conclusion, but for consistency we report the analysis that included the confounder. The results are shown as Model 3 in Table 2,  $F(2, 3397) = 151.524, p < .001$ .

The coefficient for unbalanced BR sequences ending with the less common event type was 1.482 and significant confirming that the BR manipulation had an effect on forecasts. The random effect was not significant. The adjusted odds of forecasting a flip for the balanced BR sequences were 0.35 (making the adjusted odds of forecasting a repeat 2.82). For the unbalanced BR sequences, however, the adjusted odds of forecasting a flip were 1.56 (a reversal from the 0.35 observed in the balanced BR sequences). These results demonstrate that manipulating the BR of streaky sequences in a manner that makes available the same amount of information as does the AR, but that favors a different forecast, can offset, and actually reverse, the influence of the AR on forecasts.

**Question 2.** Will the three types of TPs influence forecasts, and if so, what are their relative contributions to these forecasts?

To address this question we again used responses obtained under Conditions 1 through 6 (balanced BR, 5,100 forecasts in total). We examined three models involving TPs. The first was based on our speculation that if people are sensitive to TPs, then how often the final observation (single event, pair, or triple) appears in a sequence might be related to how much subjective weight they place on it when making their forecasts. This model had a single predictor that was the weighted average of the three types of TPs from a sequence, each weighted by the number of times its final observation occurred within the sequence (see Table 1). This model fit the data reasonably well,  $aROC = .796$  (95% CI [.784, .808]),  $r_{obs,prd} = .512$  (95% CI [.491, .532]) demonstrating that forecasts were associated with TPs.

The second model also had a single predictor; each TP was weighted accordingly by the percentage of information contributed by the single events, pairs, and triples occurring in the sequence. This model also fit reasonably well,  $aROC = .791$  (95% CI [.779, .803]),  $r_{obs,prd} = .506$  (95% CI [.485, .526]) and was not significantly different from the previous one. Note that both weighted averaging models fit the data significantly better than did Model 1 that included only AR (as indicated by the nonoverlap of the confidence intervals for their fit statistics).

In the third model, forecasts were regressed onto the three types of TPs to determine their relative influence. Because logistic regression models do not provide standardized coefficients, the inspection of which is useful in determining the relative influence of predictor variables, TPs were standardized prior to their inclusion in the analysis. The results are shown as Model 4 in Table 2,  $F(3, 5096) = 100.845, p < .001$ .

As this analysis was exploratory, we first note that the model fit the data reasonably well,  $aROC = .821$  (95% CI [.810, .832]),  $r_{obs,prd} = .558$  (95% CI [.539, .577]), significantly better than the two weighted average models. The coefficients ( $bs$ ) for TP3 and TP4 were positive and significant suggesting that subjects were sensitive to between-sequence differences in these TPs when making their forecasts. The random effects were also significant meaning that the slopes of these TPs varied significantly across subjects.

TP3 had the largest coefficient ( $b = 0.870$ , 95% CI [.736, 1.005]) indicating that forecasts were more strongly associated with differences in TP3s than with differences in TP2s ( $b = 0.075$ , 95% CI [−0.010, 0.161]) or TP4s ( $b = 0.281$ , 95% CI [0.174, 0.388]). Based on the random effects, 98% of subjects were estimated to have TP3 coefficients greater than zero, and 85% to have TP4 coefficients greater than zero. So for most subjects, as the higher-order TPs increased so did the odds of forecasting a flip

(i.e., that the 21st event in a sequence would be the opposite of the 20th).

The four models above were rerun on the total sample, and separately on the “kappa failers.” These results are provided as [online supplemental materials](#) but a few comments are provided here. In the separate analyses on the kappa failers (see Table S1 in the online supplemental materials), all model fit statistics were significantly worse and all regression coefficients were much smaller (as expected based on the lower signal-to-noise ratio in their data). In Model 1 the coefficient for AR was half that reported above and the random effect was significant (indicating greater heterogeneity). In Model 2 the coefficient for End in MC was smaller by a factor of 10 and not significant, so Hypothesis 2 was not supported. In Model 3 the coefficient for End in LC was half that reported above and the random effect was similar. In Model 4 the coefficient for TP3 (the influence of the preceding pairs) was smaller by half but significant; the coefficients for TP2 and TP4 were not significant. The random effects for these three slopes were also smaller (indicating less heterogeneity). Taken together, these supplemental analyses indicate that these subjects were using the statistical characteristics of the sequences to make their forecasts albeit to a lesser degree, they were less consistent (at the individual level), and as a group their responses contained more noise. When data from the total sample were analyzed (see Table S2 in the online supplemental materials), the substantive conclusions did not change from the primary analyses reported above (regression coefficients that were significant, although smaller, remained significant owing to the increase in sample size).

## Discussion

As one anonymous reviewer pointed out, there are no correct answers against which the forecasts may be compared to determine whether people are successful. Although this is true, it did not prevent us from uncovering which characteristics of the sequences that people pick up on when making their forecasts. This is, in part, why we were so concerned with the internal consistency of each subject and applied the kappa test to their responses. Without correct answers it is not possible to assess the validity (accuracy) of a subject's responses, but our design did permit us to assess the reliability of their responses and then explore the sequence characteristics that predicted these responses.

The results from Model 1 show that over the range of moderately streaky to moderately flippy ARs, subjects were sensitive to differences in the AR of intact sequences and adjusted their forecasts accordingly. For moderately flippy sequences ( $AR = .684$ ), the odds of forecasting a flip were 2.2 times greater than for random sequences with an AR of .5. This result is particularly interesting when considered in the context of common findings from perceived randomness studies. If our subjects had perceived such moderately flippy sequences as random, then it could be argued that the odds of forecasting a flip should be 1.

A nice feature of logistic regression models that have only a single quantitative predictor is that the negative ratio of the intercept to the slope,  $-(a/b)$ , provides an estimate of the inflection point (i.e., the value of the predictor that yields a odds of 1). From Model 1 we have  $-(-0.105/4.318) = 0.024$ , and because AR was mean-centered, we add .5 to this value to get the AR value at the inflection point,  $0.024 + .5 = .524$ . Based on this estimate our

subjects showed a very slight bias in their forecasts, but this is nothing like the values of .6 to .7 commonly seen in the perceived randomness literature.

The results from Model 2 show that nearly doubling the frequency of one event type in intact AR-neutral sequences alters forecasts in a predictable manner. Ending unbalanced BR sequences with the less common event type increased the odds of forecasting a flip by 5.84 times compared with the balanced BR sequences. Unbalanced BR sequences ending with the more common event type increased the odds of forecasting a repeat by 1.36 times compared with the balanced BR sequences.

The reason for the difference in effect sizes is not clear. A probability matching strategy cannot explain it because in both situations subjects forecasted the more frequent event type. Nor is it an instance of the gamblers' fallacy in the classic sense, because none of the sequences ended in a streak of three-of-a-kind. This asymmetry is interesting in the context of findings from work on perceived randomness. If subjects perceived neutral-AR sequences as streaky, as the randomness literature reports, then one would expect them to show asymmetry in the opposite direction; the *OR* for forecasting a repeat should have been larger than the *OR* for forecasting a flip.

The results from Model 3 show that the influence of moderate streakiness on forecasts can be reversed by providing an equal amount of conflicting information via the BR. We could not examine this effect in moderately flippy sequences because in the universe of sequences with BR = .65 and ARs of .632 or .684 ( $N = 8,184$ ), there are no sequences that conform to our design constraints ( $MRL \leq 3$ , do not end in a three-event streak, and some of which end in two of the same event type).

The results of Model 4 show that people are indeed sensitive to between-sequence differences in TPs when making forecasts. Interestingly, forecasts were most strongly associated with differences in TP3s. In other words, the relative frequency with which each event type followed a *pair of events* within a sequence appears to be the most salient of these three cues when making forecasts.

In summary, the analyses (Models 1 through 4) showed that people are sensitive to the BR, AR, and TPs of intact sequences when asked to forecast what the next event in a sequence is most likely to be. Because the number of sequences ending in two events of the same type was not uniform across conditions (recall that we added this confounder to Models 2 and 3), we wanted to correct the imbalance in the design and rerun the experiment. Experiment 2 corrected for this imbalance.

## Experiment 2

### Method

**Subjects.** One hundred twelve students, faculty, and staff on two campuses of a major university located in the southeastern United States volunteered to participate in this experiment. None participated in Experiment 1.

**Materials.** With one exception, the materials employed in Experiment 2 were identical to those used in Experiment 1. Ten sequences (and their compliments) were replaced so that the number of sequences ending in two of the same event type was uniform

across the 10 experimental conditions (four ended in either XX or OO, six ended in either XO or OX).

**Procedure.** The procedure in this experiment was identical to that described for Experiment 1. The same extensive counterbalancing was used.

**Analytic approach.** The approach used in Experiment 1 was used in this experiment.

### Results

**Data quality.** Data were obtained from 112 subjects. The average time to complete the task was 40 min. As in Experiment 1, responses were screened by checking the frequencies of Xs and Os and response alternation rates. Data from three subjects were discarded because these individuals apparently did not take the task seriously. One was excluded because he indiscriminately wrote alternating Xs and Os throughout the 100 sequences. One was excluded because although her data for the first four blocks were within limits (based on the binomial), the alternation rate of her responses exceeded .95 in the last block; she simply marked XOXOX . . . over the last 20 sequences. The third was excluded because she only wrote eight Xs (within the first 12 sequences), but 88 Os from that point on. This left data from 109 individuals for analysis. Most were students, and the majority were female (77.8%). The mean age was 27.7 years ( $SD = 8.4$ ).

As in Experiment 1, to assess warm-up and fatigue effects we compared the frequencies of X responses and the alternation rate of responses over the five consecutive blocks of 20 sequences. The average number of X responses per block ranged from 9.85 to 10.21 and the average alternation rate in responses ranged from .548 to .575. Based on these analyses warm-up and fatigue effects were not present in the data. Again we examined the frequencies of X responses and the alternation rate of responses for each set of 20 sequences to address our assumption of interchangeability among the sets. The average number of times subjects responded X within a set ranged from 9.76 to 10.40 and the average alternation rate in responses across sets ranged from .532 to .575. Based on these analyses we again felt that it was appropriate to combine responses from the five sets for substantive analyses.

As in Experiment 1, the final step in assessing the quality of our data involved calculating Cohen's kappa on the responses of each subject over the 50 complementary pairs of sequences. Fifty-seven subjects had kappa values significantly higher than chance and were retained for primary analysis. Data from the remaining 52 subjects are analyzed subsequently.

**Substantive analysis.** The same hypotheses and research questions from Experiment 1 were addressed using data from this experiment. The results are shown in Table 3.

**Hypothesis 1.** The odds of forecasting a flip will increase as AR increases from .5 through .684, and will decrease as AR decreases from .5 through .316. This was tested using 60 responses per subject under Conditions 1 through Condition 6 (3,420 forecasts in total). The results are shown as Model 1 in Table 3,  $F(1, 3418) = 178.216, p < .001$ .

In Model 1 the intercept was  $-0.099$  and the slope for AR was  $5.095$  and was significant, providing support for Hypothesis 1. The random effect for this slope was significant (100% were estimated to have a coefficient greater than zero). The estimated odds of forecasting a flip for a truly random sequence ( $AR = .5$ ) were

Table 3

*Fixed Effect (Top) and Random Effect (Bottom) Estimates for Models Predicting Forecast Decisions (Experiment 2)*

Model description	Fixed effects					Model fit statistics	
	Term	Coefficient	SE	<i>t</i>	<i>p</i>	aROC	<i>r</i> <sub>obs,prd</sub>
1. Influence of AR when BR is balanced Conditions 1–6	Intercept	−0.099	0.098	−1.020	.308	.755	.442
	AR	5.095	0.382	13.350	.001	[.739, .771]	[.415, .469]
2. Influence of BR when AR is neutral Conditions 3, 4, 9, and 10	Intercept	−0.175	0.104	−1.682	.093	.776	.483
	End in MC	−0.514	0.119	−4.321	.001	[.757, .795]	[.451, .514]
	End in LC	1.430	0.200	7.157	.001		
3. Influence of BR when in opposition to AR Conditions 1, 2, 7, and 8	Intercept	−0.875	0.119	−7.367	.001	.745	.425
	End in LC	0.890	0.140	6.346	.001	[.725, .765]	[.391, .458]
4. Relative influence of SCPs All 10 Conditions (predictors standardized)	Intercept	−0.022	0.092	−0.245	.807	.829	.569
	TP2	0.194	0.048	4.036	.001	[.818, .839]	[.551, .586]
	TP3	0.841	0.084	10.037	.001		
	TP4	0.273	0.067	4.094	.001		
	Random effects						
	Term	Variance	SE	<i>z</i>	<i>p</i>		
1. Influence of AR when BR is balanced Conditions 1–6	Intercept	0.461	0.108	4.277	.001		
	AR	3.311	1.635	2.025	.043		
2. Influence of BR when AR is neutral Conditions 3, 4, 9, and 10	Intercept	0.399	0.111	3.597	.001		
	End in MC	0.105	0.144	0.728	.466		
	End in LC	1.365	0.419	3.259	.001		
3. Influence of BR when in opposition to AR Conditions 1, 2, 7, and 8	Intercept	0.217	0.108	2.009	.044		
	End in LC	0.320	0.108	2.965	.003		
4. Relative influence of SCPs All 10 Conditions (predictors standardized)	Intercept	0.421	0.093	4.523	.001		
	TP2	0.053	0.025	2.146	.032		
	TP3	0.263	0.074	3.568	.001		
	TP4	0.151	0.048	3.121	.002		

*Note.* AR = alternation rate; aROC = area under the receiver operating characteristic curve; BR = base rate; MC = more common event type; LC = less common event type; TP = transitional probability. 95% CIs are shown in brackets.

0.91, replicating that people are equally likely to forecast a flip or a repeat when faced with truly random sequences. The estimated odds of forecasting a flip for sequences with an AR of .684 were 2.31, and for those with an AR of .316 these odds were 0.35 (the estimated odds of forecasting a repeat for these sequences were 2.82). For moderately flippy sequences the odds of forecasting a flip are 2.54 times greater than they are for a truly random sequence ( $OR = 2.31/0.91 = 2.54$ ), and for moderately streaky sequences the odds of forecasting a repeat are 2.60 times greater than they are for a truly random sequence ( $OR = 0.91/0.35 = 2.60$ ). The AR at the inflection point where the odds of forecasting a flip or a repeat are equal was .519.

**Hypothesis 2.** Among AR-neutral sequences, an unbalanced BR will influence forecasts. Specifically, unbalanced BR sequences that end in the less common event type will increase the odds of forecasting a flip, while BR sequences that end with the more common event type will increase the odds of forecasting a repeat.

This hypothesis was tested using 40 responses per subject under Conditions 3, 4, 9, and 10 (2,280 forecasts in total). The model testing our hypothesis is shown in Table 3 as Model 2,  $F(2, 2277) = 40.861, p < .001$ . Both coefficients were significant. The coefficient for unbalanced sequences ending with the less common event type was 1.430 and the coefficient for those ending with the more common event type was −0.514, providing support for

Hypothesis 2. The random effect for ending unbalanced sequences with the less common event type was significant (indicating that the effect varied across subjects, 89% were estimated to have a coefficient greater than zero) but nonsignificant for ending such sequences with the more common event.

The estimated odds of forecasting a flip were 0.84 for the balanced BR sequences. For unbalanced BR sequences ending with the less common event type, the estimated odds were 3.51, and for unbalanced BR sequences ending with the more common event type the estimated odds were 0.50. Using the balanced BR sequences as a reference, subjects were 4.17 times more likely to forecast a flip when the BR manipulation favored this forecast ( $OR = 3.51/0.84 = 4.17$ ). Subjects were 1.68 times more likely to forecast a repeat when the BR manipulation favored this forecast ( $OR = 0.84/0.50 = 1.68$ ). Taken together, these results again demonstrate that people are sensitive to changes in the BR of AR-neutral sequences and can use this cue as predicted when making their forecasts. The asymmetry in effect sizes seen in Experiment 1 remained, although here it is less dramatic.

**Question 1.** Among moderately streaky sequences where the tendency is to forecast a repeat, will manipulating the BR to favor forecasting a flip negate or maybe even reverse this tendency?

This question was addressed using 40 responses per subject in Conditions 1, 2, 7, and 8 (2,280 forecasts in total). The results are shown as Model 3 in Table 3,  $F(1, 2278) = 40.274, p < .001$ . The



coefficient for unbalanced BR sequences ending with the less common event type was 0.890 and significant confirming that the BR manipulation had an effect on forecasts. The random effect was also significant indicating that the strength of the effect varied across subjects (94% were estimated to have coefficients greater than zero).

The estimated odds of forecasting a flip for the balanced BR sequences were 0.42 (making the odds of forecasting a repeat 2.40). For the unbalanced BR sequences the estimated odds of forecasting a flip were 1.02. (a negation, but not what we would consider a reversal). Given this discrepancy with the results from Experiment 1, we reran this analysis including a dummy variable for sequences ending in two of the same event type, but the result was the same (the adjusted odds were 1.03).

**Question 2.** Will the three TPs influence forecasts, and if so, what are their relative contributions to these forecasts?

To address this question we used responses obtained under all 10 conditions (5,700 forecasts in total). This seemed justified because the number of sequences ending in two of the same event type was now uniform across all conditions. The results are shown as Model 4 in Table 3,  $F(3, 5696) = 58.277, p < .001$ .

The model fit the data reasonably well,  $aROC = .829$  (95% CI [.818, .839]),  $r_{obs,prd} = .569$  (95% CI [.551, .586]). The coefficients for all three types of TP were positive and significant suggesting that subjects were sensitive to all three TPs as they made their forecasts. The random effects were also significant, meaning that the influences of the TPs varied significantly across subjects.

TP3 (based on previous pairs) again had the largest coefficient ( $b = 0.841$ , 95% CI [0.677, 1.005]) indicating that forecasts were more strongly associated with differences in TP3s than with differences in TP2s ( $b = 0.194$ , 95% CI [0.100, 0.289]) or TP4s ( $b = 0.273$ , 95% CI [0.142, 0.404]). As the confidence intervals for the TP2 and the TP4 coefficients overlap, we cannot say that one or the other of these made the second strongest contribution. Based on the random effect, 80% of subjects were estimated to have TP2 coefficients greater than zero, 95% to have TP3 coefficients greater than zero, and 76% to have TP4 coefficients greater than zero.

Given the significant random effects for all three TPs, we decided to run single-subject logistic regression models on the 100 forecasts from each subject to better understand the relative importance of the TP cues at the individual level. Initially we included all three TPs in each regression model, but many subject's data showed suppression effects that can distort conclusions (see Beckstead, 2012) and so we instead report the results from a forward-selection approach.

Fifty-five of the 57 subjects (96%) had at least one type of TP as a significant predictor of their forecasts. Thirty-seven subjects had a single significant TP predictor, 17 had two significant TP predictors, and for one individual all three TPs were significant. Among the 37 individuals with a single significant predictor, it was TP3 for 25 of them, TP4 for eight, and TP2 for four. Among the 17 subjects with two significant predictors, TP3 was the first to enter for 12 of individuals and the second to enter for five. TP2 was a significant predictor for five individuals, and TP4 was a significant predictor for six. In summary, these single-subject regressions revealed that TP3 was used by 76% of the subjects, TP4 by 38%, and TP2 by 20% as they made their forecasts.

The four models presented above were rerun on the total sample, and separately on the "kappa failers." These results are provided as supplemental materials (see Tables S3 and S4 in the online supplemental materials). Similar to Experiment 1, in the separate analyses on the kappa failers, all model fit statistics were significantly worse and all regression coefficients were much smaller (as expected based on the lower signal-to-noise ratio in these data). When data from the total sample were analyzed, the substantive conclusions did not change from the primary analyses reported above (i.e., regression coefficients that were significant, although smaller, remained significant owing to the increase in sample size).

## Discussion

This experiment replicated Experiment 1 with one important difference, this being that in the current experiment the number of sequences that ended in two events of the same type was uniform across all 10 conditions. The findings were quite consistent with those from Experiment 1, but for a few details highlighted below.

As in Experiment 1, Model 1 showed that subjects forecast flips when sequences are moderately flippy and forecast repeats when sequences are moderately streaky. The model parameters indicate that forecasts of flips and repeats are equally likely when a sequence is truly random (i.e.,  $AR = .5$ ). The AR value at which the odds of forecasting a flip or a repeat are equal was .519, quite similar to that found in Experiment 1.

The results from Model 2 demonstrated that manipulating the BR of AR-neutral sequences can induce subjects to exhibit a form of probability matching, predicting the more common event whether sequences ended in the more common, or the less common, event type. This effect was again found to be asymmetric although less so than in Experiment 1.

Model 3 addressed the question of how forecast tendencies would be influenced when sequences contained similar amounts of conflicting information made available from the AR and the BR. Similar to Experiment 1 this manipulation had a significant effect on forecasts, however, by comparison the effect was weaker, negating, but not reversing the influence of AR.

Model 4 examined the influence of the three types of TPs on forecasts made under all 10 conditions. As in Experiment 1, TP3s appeared to be a key characteristic of sequences that influenced forecasts. It could be argued, however, that subjects were simply attending to the last two events in the sequences rather than to between-sequence differences in TP3s. To address this possibility, we fit three supplemental models for comparison. The first used TP3 values as the only predictor, the second used the dummy variable for sequences that ended with two of the same event type, and the third model used the actual values of events 19 and 20 (X coded 1 and O coded 0) in the sequences. The model with only TP3 fit the data reasonably well,  $aROC = .808$  (95% CI [.797, .819]),  $r_{obs,prd} = .535$  (95% CI [.516, .553]). The other two models fit significantly worse than this one based on their fit indices and confidence intervals. The fit of the model with the dummy variable for sequences ending with two of the same event type was  $aROC = .720$  (95% CI [.706, .733]),  $r_{obs,prd} = .386$  (95% CI [.364, .408]), and the fit of the model using the values of 19th and 20th events was  $aROC = .641$  (95% CI [.627, .655]),  $r_{obs,prd} = .260$  (95% CI [.236, .284]). These supplemental analyses offer indirect support to the proposition that subjects were attending to TP3s.

Subsequent single-subject logistic regression analyses were also conducted to assess the influence of the three types of TPs. These analyses confirmed that for the majority of subjects, transitional probabilities associated with previous pairs were the dominant cue used in their forecasts.

Both Experiments 1 and 2 demonstrated that forecasts of what comes next are influenced by between-sequence differences in AR and differences in higher-order TPs; the influence of TP3s was notably the strongest. What was not addressed in these experiments, however, is whether these statistical characteristics explicitly influence peoples' *perceptions* of sequences. This topic is the focus of Experiment 3.

### Experiment 3

The purpose of this experiment is to assess the perceived "forecastability" of various binary sequences. The paired-comparison method was employed because it is particularly useful for quantifying perceived differences when the type of judgment being made is rather abstract and where differences among the objects being judged are subtle in nature. In this experiment subjects were presented with 36 pairs of sequences that differed in their statistical characteristics and asked to choose the sequence within each pair that makes it easier to predict what comes next.

#### Method

**Subjects.** One hundred one students at two campuses of a major university located in the southeastern United States volunteered to participate in this experiment. None participated in the previous experiments.

**Materials.** Nine binary sequences, each of length 20, were created using the symbols "X" and "O." All sequences contained 10 Xs and 10 Os. These sequences are shown, along with their alternation rate and relevant transitional probabilities, in Table 4.

Sequence A, which appeared in the task instructions, was included in the stimulus set to establish a minimum standard for attention to the task. When paired with any of the other eight

sequences, A should always be chosen as the sequence that makes it easier to predict what comes next (because it is simply alternating Xs and Os). Therefore, any subject not choosing sequence A each time it appeared in a pair would be considered "insufficiently engaged" in the task and would be dropped from further analysis.

Excluding sequence A, the remaining eight sequences (B through I) were selected to create a set that varied in forecastability from multiple points of view (see predictions below). Two sequences share one of four AR values (B, C = .368; D, E = .474; F, G = .526; and H, I = .632), and correspondingly, one of four TP2 values (.333, .444, .556, .667, respectively). Within each of these pairs the TP3 values differ.

Two sequences (B and I) have TP2 and TP3 values that are in agreement regarding whether the 21st event is likely to be the same as, or the opposite of, the 20th event; that is, both values are below, or above, .5. Four sequences (C, E, G, and H) have conflicting transitional probabilities that favor different forecasts. For example, sequence C has a TP2 of .333 but a TP3 of .600. Finally, two sequences (D and F) have a TP3 value of .500, meaning that based on the final pair of events, the 21st event is equally likely to be the same as, or the opposite of, the 20th. In other words, for these two sequences TP3 provides no clue as to what comes next.

The nine sequences were presented in a paired-comparison format (a series of 36 pairs in all). Pairs were arranged for presentation according to the method of balanced optimal order developed by Ross (1934). The method is designed to avoid regular repetitions that might influence judgment by maintaining the greatest possible spacing between pairs involving any given member of a stimulus set, and by balancing the number of times any given stimulus appears as the first and second member of a pair. In the current experiment, pairs involving the same sequence were separated by a maximum of five and a minimum of four positions within the series. Each sequence appeared eight times (four as the first, and four as the second, member of a pair). A second version of the questionnaire was prepared in which the series of pairs was presented in reverse order to control for possible fatigue effects. Subjects were provided with the following instructions:

Look at this sequence of 20 symbols: OXOXOXOXOXOXOXOXOXOX. If you had to predict what the 21st symbol is most likely to be, you would probably say "O" because of the obvious pattern in the sequence. In comparison, this sequence XXOXOOXXOOXXOOXXOX has a less obvious pattern and so it's more difficult to predict what the 21st symbol is most likely to be. In this experiment you will be presented with *pairs* of sequences to consider. We want you to tell us which sequence in each pair that you think makes it easier to predict the 21st symbol.

For example, if you are given this pair of sequences:

1. OXOXOXOXOXOXOXOXOXOX 2. XXOXOOXXOOXXOXOXOX

you would choose number 1 because it is obviously easier to predict the 21st symbol in this sequence than it would be with sequence number 2.

On the next 2 pages there are 36 pairs of sequences for you to examine. **Your task is to choose the sequence in each pair that you think makes it easier to predict the 21st symbol.** You DO NOT have to actually predict the 21st symbol, just choose the sequence that

Table 4  
Sequences Presented in Experiment 3: Paired-Comparison Task

ID	Stimulus sequence	AR	TP2	TP3	Scale value
A	OXOXOXOXOXOXOXOXOXOX <sup>a</sup>				
B	XXXXOXXXXXOOOXXOXOXO	.368	.333	.400	47.9
C	OXXXXOXXXXXOXOXOXOXOX	.368	.333	.600	46.5
D	OXOOOXOOOXOXOXOXOXOX	.474	.444	.500	42.2
E	XXOOXXOOOXOXOXOXOXOX	.474	.444	.750	59.5
F	XXOOXOOXOXOXOXOXOXOX	.526	.556	.500	53.4
G	OXXOOXOXOXOXOXOXOXOX	.526	.556	.250	57.9
H	XOXXOXOOOXOXOXOXOXOX	.632	.667	.400	52.7
I	OXXXOXOXOXOXOXOXOXOX	.632	.667	.600	39.9

*Note.* N = 91. AR = alternation rate; TP2 = second-order transitional probability (that the 21st event will be the opposite of the 20th) based on previous single events; TP3 = third-order transitional probability (that the 21st event will be the opposite of the 20th) based on previous pairs of events. Scale value = perceived ease of forecasting the 21st event. Critical range = 1.44; any pair of sequences differing by this amount were perceived as significantly different from one another ( $p < .05$ ).

<sup>a</sup>Control stimulus included to establish a minimum standard for task engagement.

you think makes it easier to do so. Do not over think your choice; go with your first impression. There are no right or wrong answers, but please choose one sequence in each pair.

**Procedure.** Data were collected in classroom settings. Subjects were randomly assigned to complete one of the two forms of the questionnaire.

**Three predictions.** (a) If people respond primarily to differences in AR as suggested by perceived randomness research, then sequences *H* and *I*, with  $AR = .632$ , will be perceived as random and therefore be judged as the least forecastable. As the AR decreases toward .368, (and sequences presumably are perceived as becoming less random) perceived forecastability should increase. (b) If people respond primarily to differences in TP3s, then as TP3 values increasingly deviate (absolutely) from .5, perceived forecastability should increase. Sequences *E* and *G* should be judged as the most forecastable. (c) If people respond equally to TP2s and TP3s, then because sequences *B* and *I* are the only two that have transitional probabilities in agreement regarding the 21st event, these two sequences should be judged as the most forecastable.

## Results

**Data quality check.** Of the total  $N = 101$  who completed all 36 pairs, 10 failed the basic test of sufficient engagement by not choosing sequence *A* in all of its pairings with other sequences. This left  $N = 91$  for subsequent analysis. The average age was 31.0 years ( $SD = 7.8$ ), and 74 (80%) were female. There were no differences in the results between the two orders of presentation. The task took approximately 10 min to complete.

**Scaling perceived forecastability.** After excluding sequence *A*, the number of times that each sequence (*B* through *I*) was chosen as making it easier to predict the 21st event was tabulated for each subject (range of possible scores for a sequence 0 to 7). These scores were then aggregated across subjects to scale the sequences on ease of forecastability using rank-sum scaling (Dunn-Rankin & King, 1969). The basic assumption of the method is that the scale values are proportional to the sum of the scores obtained from the paired-comparison procedure. The maximum and the minimum possible rank totals act as a convenient and interpretive frame of reference within which the sequences are scaled. This scaling method has units that are equal in a variance stable sense, meaning that a specific difference between the rank totals (and hence the scaled values) has the same probability of occurrence wherever the rank totals may be located within the range. This variance stability allows for the calculation of a critical range value (analogous to the Tukey method of multiple comparisons among means) providing the opportunity to test for significant differences between pairs of sequences. The scale values for each sequence are shown in Table 4.

The rank ordering of the sequences on the scale of perceived forecastability was: *E, G, F, H, B, C, D, I*. The critical range was 1.44, meaning that any two sequences that differed by this amount can be considered as significantly different ( $p < .05$ ) on the scale. Sequences *E* and *G* (that had the most extreme TP3 values) had the highest scale values (indicating that they were perceived as the most forecastable) and these values were significantly higher than those of the remaining six sequences. The eight scale values correlated .645 with the absolute deviations of the TP3 values from

.5, whereas their correlation with AR values was only .002. Although not considered when selecting the sequences, we calculated the amount of information provided by the preceding pairs within each sequence and also correlated these values with scale values ( $r = .640$ ).

## Discussion

This experiment demonstrated that comparative judgments of forecastability are related to between-sequence differences in TP3s. In other words, when judging forecastability, people attend to the relative frequencies with which a symbol follows preceding pairs of symbols in a sequence. Sequences with TP3 values that deviated the most from .5 were judged as being the most easy to forecast. Perceived forecastability was uncorrelated with between-sequence differences in AR. It would seem that people attend to different statistical characteristics of sequences when thinking about forecasting than when thinking about randomness.

Although the results regarding the influence of TP3s corroborate those from Experiments 1 and 2, generalizability must be made with some caution as only eight sequences were involved in this experiment. Nevertheless, because the results here are based on 28 comparative judgments from each subject (rather than only seven absolute judgments had a rating task been used), the results have a certain degree of credibility.

## General Discussion

In this article we have presented three experiments dealing with how people respond to the statistical characteristics of intact binary sequences. In Experiments 1 and 2 we have demonstrated that when presented with such sequences and asked to forecast the next event in each sequence, people are sensitive to differences in base rate, alternation rate and higher-order sequential dependencies (i.e., transitional probabilities associated with preceding pairs and triples) and can use these characteristics as implicit cues, adjusting their forecasts accordingly. We began by examining AR and BR separately as these have been studied more extensively than higher-order sequential dependencies (e.g., TP3s and TP4s).

People responded to moderately streaky sequences with a tendency to forecast repeats, but rather than treating moderately flippy sequences as random, they displayed a tendency to forecast that the next event will be the opposite of the last. In both experiments Model 1 provided an estimate of the AR at which the odds of forecasting a flip or a repeat are equal (.524 in Experiment 1, .519 in Experiment 2). People showed a very slight bias in their forecasts, but this is nothing like the values of .6 to .7 commonly seen in the perceived randomness literature (e.g., Falk, 1981; Falk & Konold, 1997; McDonald, 2009; Scholl & Greifeneder, 2011). Perhaps when examining intact sequences for the purposes of detecting a pattern and making a forecast, rather than judging the randomness of a sequence, people have a different threshold for randomness.

When faced with sequences in which the AR provides little or no information about what comes next, people engaged in a form of probability matching, relying on the BR of the two event types to make their forecasts. Unbalanced BR sequences ending with the less common event type increased the odds of forecasting a flip (i.e., forecasting the more common event), whereas sequences

ending with the more common event type increased the odds of forecasting a repeat (again, forecasting the more common event). In both experiments the effect was asymmetric; ending sequences with the less common event type produced a stronger effect. One possible explanation for the asymmetry may be that the TPs for the sequences in the two unbalanced BR conditions were themselves unbalanced. Post hoc inspection revealed that sequences ending with the less common event type had, on average, more extreme TP values (deviating from .5) than did sequences ending with the more common event type.

When the AR and BR of sequences were simultaneously manipulated to favor opposing forecasts, the results differed somewhat between experiments. In both experiments the influence of moderate streakiness was canceled by providing an equal amount of conflicting information via the BR. In Experiment 1 the BR manipulation actually reversed the influence of AR. Why the difference in results was obtained across the two experiments is not clear.

Higher-order transitional probabilities associated with preceding pairs (TP3s) and triples (TP4s) were shown to influence peoples' forecasts in both experiments. Forecasts were more strongly related to preceding pairs than to preceding single events whether the influence of such single events was modeled by second-order transitional probabilities or by AR.

Our task instructions (identical for both experiments) stated that the sequences had been generated by mathematical algorithms. This type of generating mechanism falls somewhere between the animate (e.g., human) and inanimate (e.g., coin flip, slot machine) sources traditionally examined in studies of perceived randomness (see Oskarsson et al., 2009 for review). Although the full impact of this detail is unclear, it may have steered people away from both positive-recency and the negative-recency biases, meaning that the number of flip and repeat forecasts were about equal. In the first Experiment 4,405 of 8,500 forecasts (51.8%, 95% CI [50.8%, 52.9%]), were flips, and 2,829 of 5,700 forecasts (49.6%, 95% CI [48.3%, 50.9%]) were flips in the second experiment. The instructions also mentioned that we were interested in understanding how people are able to distinguish randomness from subtle patterns, or regularities in sequences, but gave no indication as to what such patterns might be or what to look for specifically. What people appear to find is third-order redundancy.

When people inspect a sequence with the intent of forecasting what comes next, they may perceive the sequence differently than they do when their motive is to assess its randomness. We addressed this possibility in Experiment 3 that focused on perceived forecastability. As a sequence's forecastability seemed to us as perhaps even more nebulous to judge than its randomness (if that's possible), we used the method of paired comparisons. When judging forecastability, people attended to the relative frequencies with which a symbol follows preceding pairs of symbols (i.e., TP3s) in a sequence but not to AR. This suggests that the constructs of perceived randomness and perceived forecastability are not simply the converse of one another; if they were, then one would expect perceived forecastability to be negatively correlated with AR over the range examined.

Although not investigated in the current article, higher-order TPs may play a role in perceived randomness. In their seminal work, Falk and Konold (1997) wrote that the general problem of defining and measuring randomness remained unresolved because

of the limitation of considering only second-order conditional uncertainty (i.e., AR) and ignoring redundancies of higher orders. Since then much of the work in this arena has continued to focus on AR. Our interest in the distribution of higher-order TPs may help to address this limitation.

In the universe of 20-event binary sequences with BR = .5, the ARs as close as possible to .5 are .474 and .526. There are 15,876 of each AR (31,752 sequences in total). They all contain the same trivial amount of information associated with previous single events (0.004 bits). Yet, for all these sequences, the amount of information from previous pairs is greater than that from previous singles (range for pairs: 0.012 to 0.996 bits). For 31,239 sequences the amount of information from previous triples is greater than that from previous singles (range for triples: 0.005 to 0.984 bits). Notably, there are no sequences where the amount of information from previous singles is greater than that from pairs and greater than that from triples.

The distribution of the TP3s (from pairs) and TP4s (from triples) predicting that the 21st event will be the opposite of the final vary across the 31,752 sequences with values of .000, .250, .333, .500, .667, .750, and 1.000. There are 3,828 sequences that have TP3 = TP4 = .500, meaning that their final pairs and triples provide no clues when forecasting what the next event might be. We consider this small subset (3,828 of 524,288) to be the "most random" sequences in this universe. It would be interesting to see how people judge the randomness of these sequences in comparison to others with the same ARs.

Investigators who use intact binary sequences as stimuli should be cautious and not assume that because a set of sequences has the same AR that they are interchangeable. For example, in their Experiment 1, Scholl and Greifeneder (2011) presented 36 unique sequences of 21 events (and complements, 72 in total) and asked subjects to rate the randomness of each sequence. There were three AR conditions (.4, .5, and .6) crossed with three MRL conditions (3, 4, and 5) in their experiment. A key finding, shown in their Figure 2, was that MRL showed an impact on judgments, but only for the AR = .4 condition. The authors provide nine unique sequences illustrating the experimental conditions in their Table 1. We calculated the TP3 distribution and used these to calculate the amount of information contributed by pairs in these example sequences. The average amount of information from pairs is much greater for the three .4 AR sequences (0.36 bits) than for the three .5 AR sequences (0.07 bits) and the three .6 AR sequences (0.08 bits). We cannot say that this imbalance is responsible for their finding because we do not know the information values for the remaining 27 unique sequences, but if these other sequences show the same pattern across conditions it calls into question the importance of MRL as a moderator of AR.

Comparing our results to the early work on probability learning that involved trial-by-trial forecasts is tricky, because in that paradigm each forecast is followed by feedback. This meant that forecasts were related not only to the sequential dependencies in the series of stimuli presented, but also to the preceding forecasts made. To our knowledge there were two studies that investigated forecast behavior in a two-choice probabilistic task in which the independent variable is the conditional probability of the stimuli in the sequence. The first of these was conducted by Hake and Hyman (1953) and the second by Anderson (1960). We discuss each in turn.



Hake and Hyman (1953) were the first to investigate people's ability to learn sequential conditional probabilities. In their experiment subjects experienced a predetermined sequence of 240 binary symbols (a vertical column or a horizontal row of four small neon bulbs). Immediately following a warning signal, and before the symbol was presented, subjects were required to forecast which of the two possible symbols would occur on that trial. The data analyzed were the series of forecasts made by each subject. The focus of the analysis was to show the degree to which each forecast (except for the first) was determined by the preceding events (both preceding stimuli and preceding forecasts). They examined not only the influence of the preceding single event, but also that of the preceding pair and preceding triple. In separate analyses using information theory measures (entropy), Hake and Hyman showed that forecasts were dependent on preceding stimuli and to a lesser degree on preceding forecasts. They report that the preceding stimulus, and the preceding pairs of stimuli influenced forecasts, but the preceding stimulus triple had no appreciable influence on forecasts. Furthermore, their analyses indicate that the influence of the preceding pair of stimuli was stronger than that of the preceding single stimulus.

An important distinction between their data and ours is that with trial-by-trial feedback it is not possible to unambiguously disentangle the influence of the stimulus series from the influence the subjects' preceding forecasts. As they put it "the observer does not predict the future behavior of a series of events purely on the basis of the series. The observer *contaminates* his judgment by including his own previous behavior and the 'correctness' of his own previous behavior in the set of events which influences his expectancy of future events" (Hake & Hyman, 1953, p.73, [italics added]). Our Experiments 1 and 2 avoided such contamination by presenting the subject with several (relatively short) sequences with different sequential conditional probabilities, each in its entirety, requesting a forecast for each sequence, and providing no feedback following forecasts. This meant that our analyses assessed the influences of the statistical characteristics of the stimulus sequences only. Consistent with Hake and Hyman (1953) our results show that forecasts were influenced by the statistical characteristics of the stimulus sequences. As in their analyses, we found the influence of the preceding pairs to be stronger than the influence of the preceding single events. In contrast to their findings, we also found that forecasts were (to a lesser degree) influenced by the preceding triples. This may be because our design was more sensitive or in their design it was more difficult to disentangle this stimulus influence from sequential dependencies in the subject's responses.

Anderson (1960, Experiment 2) investigated the extent to which people could learn conditional probabilities using the same trial-by-trial forecast paradigm as Hake and Hyman (1953). Subjects were randomly assigned to experience one of nine sequences with different alternation rates (alternation rates ranged from .10 to .90 in steps of .10). In all nine conditions, the base rates of the two stimuli were approximately equal. Each subject experienced 300 acquisition trials (on which we focus), followed by 200 transfer trials and then 60 recovery trials. The dependent variable used in his analysis was the frequency of repetition forecasts, defined as responding with the same forecast on the current trial that had been made on the preceding trial.

To assess the impact of sequential dependencies, Anderson tabulated the proportion of repetition forecasts following preced-

ing stimulus subsequences (tuples, as he called them) of various lengths, over Trials 101 to 300. For instance, the proportion of times that repetition responses followed a run of one, two, or three stimuli of the same type was calculated. Repetition responses were also tabulated for alternating tuples of different lengths. This procedure was carried out for all tuples of length 1 to 8 and summarized in his Figure 3, Table 2, and Table 3. Careful inspection of these summaries reveals that repetitive responses were influenced more by the preceding pairs and triples than by preceding singles (or longer preceding tuples). Whether the preceding pairs or the preceding triples had the greatest impact varied from one alternation rate condition to another, but no formal comparison was provided. Although his analysis was quite different from ours, it did suggest that people's forecasts are influenced by higher order transitional probabilities. Like the results reported by Hake and Hyman (1953), however, it is not possible to unambiguously disentangle the influence of sequential dependencies in the stimulus sequence from those in subject's responses.

As in any article that attempts to make substantive claims, there are some limitations to bear in mind. First, we strove to keep the mathematical treatment of relevant concepts to a minimum in the hopes of facilitating readability; this meant, however, that rigorous proofs of various statements could not be attempted. In light of this limitation, we have provided references to the primary works for those readers who wish more detail on the derivation of formulae and ideas. Second, the universe is big, really big, and we only examined peoples' responses to 128 of the 1,048,576 possible 20-event sequences in it. It remains to be seen how well our results generalize to these other sequences. Third, in Experiments 1 and 2 nearly half of the subjects failed our kappa test of within-subject response consistency. We believe that although this feature of our experimental design strengthens internal validity by increasing the signal-to-noise ratio in the data analyzed, it may detract from external validity. The purpose of these two experiments was to demonstrate that people are capable of using the statistical characteristics of intact sequences when making forecast decisions; future work might focus on mechanisms, boundary conditions, or moderating individual difference variables such as working memory. Fourth, in Experiment 3 that examined perceived forecastability using the paired-comparison method, only eight sequences were used. Although we could have examined a larger set of sequences had we asked subjects simply to rate the perceived forecastability of each sequence, we believe that the paired-comparison method is better suited to revealing subtle perceptual distinctions, albeit on fewer objects of judgment.

Bearing these limitations in mind, we do have a certain amount of confidence in the veracity of our findings. The results of Experiment 1 are, by and large, replicated in Experiment 2. Furthermore, the extensive attention to counterbalancing in the design of these experiments strengthens this confidence. The design of Experiment 3 included a means of identifying (and removing) subjects who did not meet a minimum standard for task engagement. This measurement error reduction strategy further bolsters confidence in our findings.

In this article we have examined the complex structure of seemingly very simple stimuli: binary sequences. We coded these sequences in terms of base rate, alternation rate, and higher-order transitional probabilities associated with preceding pairs and triples. Our aim in doing so was to determine which of these

statistical characteristics are used as implicit cues when forecasting the next event in such sequences and how these characteristics affect peoples' perception of a sequence's forecastability.

Like Brunswik, we believe that detailed analysis of the stimuli presented in psychological experiments is as important as the analysis of the responses obtained. This attention to the subtle differences in the characteristics of our stimuli revealed that peoples' forecasts were more influenced by preceding pairs and triples than by preceding single events (i.e., by AR). Further, how people perceive binary sequences seems to be dependent on why they are looking at them or on what they are trying to achieve in doing so. Perceived forecastability and perceived randomness are not simply the converse of one another; examining a sequence with the goal of forecasting what comes next seems to encourage a different way of looking at it.

## References

- Alberoni, F. (1962). Contribution to the study of subjective probability. I. *Journal of General Psychology*, 66, 241–264. <http://dx.doi.org/10.1080/00221309.1962.9711840>
- Altmann, E. M., & Burns, B. D. (2005). Streak biases in decision making: Data and a memory model. *Cognitive Systems Research*, 6, 5–16. <http://dx.doi.org/10.1016/j.cogsys.2004.09.002>
- Anderson, N. H. (1960). Effect of first-order conditional probability in two-choice learning situation. *Journal of Experimental Psychology*, 59, 73–93. <http://dx.doi.org/10.1037/h0049023>
- Attneave, F. (1959). *Applications of information theory in psychology*. New York, NY: Henry Holt.
- Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, 12, 428–454. [http://dx.doi.org/10.1016/0196-8858\(91\)90029-1](http://dx.doi.org/10.1016/0196-8858(91)90029-1)
- Beckstead, J. W. (2012). Isolating and examining sources of suppression and multicollinearity in multiple linear regression. *Multivariate Behavioral Research*, 47, 224–246. <http://dx.doi.org/10.1080/00273171.2012.658331>
- Brunswik, E. (1956). *Perception and representative design of psychological experiments*. Berkeley, CA: University of California Press.
- Carlson, K. A., & Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. *Organizational Behavior and Human Decision Processes*, 104, 113–121. <http://dx.doi.org/10.1016/j.obhdp.2007.03.004>
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582. <http://dx.doi.org/10.1037/0003-066X.34.7.571>
- Dunn-Rankin, P., & King, F. J. (1969). Multiple comparisons in a simplified rank method of scaling. *Educational and Psychological Measurement*, 29, 315–329. <http://dx.doi.org/10.1177/001316446902900207>
- Falk, R. (1981). The perception of randomness. *Proceedings of the Fifth International Conference for the Psychology of Mathematics Education*, 1, 222–229.
- Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, 104, 301–318. <http://dx.doi.org/10.1037/0033-295X.104.2.301>
- Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York, NY: Wiley.
- Goodfellow, L. D. (1938). A psychological interpretation of the results of the Zenith radio experiments in telepathy. *Journal of Experimental Psychology*, 23, 601–632. <http://dx.doi.org/10.1037/h0058392>
- Green, D. R. (1982). Testing randomness. *Teaching Mathematics and Its Applications*, 1, 95–100. <http://dx.doi.org/10.1093/teamat/1.3.95>
- Hake, H. W., & Hyman, R. (1953). Perception of the statistical structure of a random series of binary symbols. *Journal of Experimental Psychology*, 45, 64–74. <http://dx.doi.org/10.1037/h0060873>
- Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education*, 24, 392–414. <http://dx.doi.org/10.2307/749150>
- Luce, R. D. (2003). Whatever happened to Information Theory in psychology? *Review of General Psychology*, 7, 183–188. <http://dx.doi.org/10.1037/1089-2680.7.2.183>
- Matthews, W. J. (2013). Relatively random: Context effects on perceived randomness and predicted outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1642–1648. <http://dx.doi.org/10.1037/a0031081>
- McDonald, F. E. J. (2009). *Understanding randomness via the perception and prediction of binary sequences* (Unpublished dissertation). University of New South Wales, Sydney.
- Miller, G. A., & Frick, F. C. (1949). Statistical behavioristics and sequences of responses. *Psychological Review*, 56, 311–324. <http://dx.doi.org/10.1037/h0060413>
- Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin*, 135, 262–285. <http://dx.doi.org/10.1037/a0014821>
- Ross, R. T. (1934). Optimum orders for the presentation of pairs in the method of paired comparisons. *Journal of Educational Psychology*, 25, 375–382. <http://dx.doi.org/10.1037/h0070754>
- Sanderson, Y. B. (2012). Color charts, esthetics, and subjective randomness. *Cognitive Science*, 36, 142–149. <http://dx.doi.org/10.1111/j.1551-6709.2011.01198.x>
- Scholl, S. G., & Greifeneder, R. (2011). Disentangling the effects of alternation rate and maximum run length on judgments of randomness. *Judgment and Decision Making*, 6, 531–541.
- Shannon, C. E. (1948a). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423. <http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shannon, C. E. (1948b). A mathematical theory of communication: Part III: Mathematical preliminaries. *The Bell System Technical Journal*, 27, 623–656. <http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x>
- Wagenaar, W. A. (1970). Appreciation of conditional probabilities in binary sequences. *Acta Psychologica*, 34, 348–356. [http://dx.doi.org/10.1016/0001-6918\(70\)90030-2](http://dx.doi.org/10.1016/0001-6918(70)90030-2)
- Wiegiersma, S. (1987). The effects of visual conspicuousness and the concept of randomness on the recognition of randomness in sequences. *Journal of General Psychology*, 114, 157–165. <http://dx.doi.org/10.1080/00221309.1987.9711066>

Received June 30, 2019

Revision received February 28, 2020

Accepted April 19, 2020 ■