Assessing the “bias” in human randomness perception

Paul A. Warren, Umberto Gostoli, George D. Farmer, Mark Boyle, Wael El-Deredy
(paul.warren@manchester.ac.uk)
School of Psychological Sciences, University of Manchester,
Manchester, M13 9PL, UK

Andrew Howes
(howesa@bham.ac.uk)
School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK

Ulrike Hahn
(u.hahn@bbk.ac.uk)
Department of Psychological Sciences
Birkbeck University of London
London, WC1E 7HX, UK

Abstract

Human randomness perception is commonly described as biased. This is because when generating random sequences humans tend to systematically under- and over-represent certain sub-sequences relative to the number expected from an unbiased random process. In a purely theoretical analysis we have previously suggested that common misperceptions of randomness may actually reflect genuine aspects of the statistical environment, once cognitive constraints are taken into account which impact on how that environment is actually experienced. In the present study we provide a preliminary test of this account, comparing human-generated against unbiased process-generated binary sequences. Crucially we apply metrics to both sets of sequences that reflect constraints on human experience. In addition, sequences are compared using statistics that are shown to be more appropriate than a standard expected value analysis. We find preliminary evidence in support of our theoretical account and challenge the notion of bias in human randomness perception.

Keywords: Randomness perception; random sequence generation, cognitive biases.

Introduction

Randomness is the flip side of statistical structure. Researchers interested in human beings as ‘intuitive statisticians’ have consequently long been interested in people’s ability to identify patterns of data as random. A long tradition of research has reached rather negative conclusions about people’s intuitive understanding of randomness. Whereas early studies focussed primarily on people’s ability to generate random sequences (see e.g., Wagenaar, 1972), later work has also examined people’s ability to judge sequences as random (see e.g., Kahneman & Tversky, 1972; Bar-Hillel & Wagenaar, 1991; Oskarsson et al. 2009).

Both studies of sequence generation and production have found evidence of similar biases, in particular a bias toward over-alternation between the different possible outcomes, such as ‘heads’ (H) or ‘tails’ (T), in binary sequences. This alternation bias has frequently been interpreted as evidence for a belief in the ‘gambler’s fallacy’ (GF), that is, the erroneous belief that an increasing run of one outcome (e.g., HHHHHH…) makes the other outcome ever more likely (but see e.g., Edwards, 1961). Such a belief, which can indeed be found among gamblers around the world (Clotfelter & Cook, 1993; Terrell, 1998; Tonneato et al., 1997; Croson & Sundali, 2005), may reflect a mistaken conception of random processes as ‘self-correcting’ in such a way as to maintain an equal balance between the possible outcomes (for other explanations see e.g., the review by Hahn, 2011).

However, the concept of randomness is a difficult, and often counter-intuitive, one not just for gamblers or experimental participants, but also for experimenters (on the concept of randomness see e.g., Beltrami, 1999), and extensive critiques have shown much of the empirical research on lay understanding of randomness to be conceptually flawed (see in particular, Ayton, Hunt & Wright, 1989; Nickerson, 2002; but also Lopes, 1982).

Aforementioned evidence from real-world gamblers aside, it is thus less clear than might be expected how good or bad lay people’s ability to both discern and mimic the output of random sources actually is.

Research with novel tasks, that do not suffer from the conceptual flaws identified, have tended to confirm some element of bias in people’s performance (e.g., Rapaport & Budescu, 1982; Olivola & Oppenheimer, 2008) while finding also that participants’ performance is considerably better than deemed by past research (see e.g., Lopes & Oden, 1981; Nickerson & Butler, 2009).

In particular, it has been argued that people’s performance may actually be quite good given their actual experience of random sequences, whether inside or outside the lab. William and Griffiths (2013) show how seemingly poor performance on randomness judgment tasks may stem from the genuine paucity of the available statistical evidence. Hahn & Warren (2009) similarly argue that common biases and misperceptions of randomness may actually reflect genuine aspects of the statistical environment, once it is taken into account how that environment is actually experienced. Specifically, Hahn and Warren demonstrate that if human experience of a stream of binary random events is assumed to be i) finite and ii) constrained by the limitations of short-term memory and/or attention, then based upon highly counter-intuitive mathematical results, not all binary sub-strings are equally likely to occur.

We describe this theoretical work next in more detail, before going on to present the results of a behavioural experiment that looks for preliminary evidence that human perception of randomness conforms to the theoretical treatment outlined.

Experiencing Random Sequences

Hahn & Warren's account relies upon a simple model of how a human might experience an unfolding sequence of random events. It is proposed that humans have a limited capacity *window of experience* of length *k* that has access to the present event and preceding *k*-1 events. This window slides one event at a time through an unfolding finite sequence of length $n > k$. That humans could only ever experience a finite stream of events is incontrovertible. Further, given the well-characterized bounds on human short-term memory capacity and/or attention span, this limited capacity, sliding window of experience account seems plausible.

Crucially, when sub-sequences of length *k* are counted amongst a longer finite sequence of length *n* using the sliding window analysis suggested above, certain sub-sequences are more likely to not occur, *even when the generation process is unbiased*. In particular perfect runs of one outcome have highest *non-occurrence probability* (or conversely lowest occurrence rate), followed by perfect alternations of the two outcomes. This highly counter-intuitive mathematical result is illustrated in figure 1B; the unbroken line represents the occurrence rates for the 16 possible subsequences of length 4. For example, the occurrence rate for the perfect run subsequence 0000 is around 0.47 meaning that this subsequence doesn't appear at all on around 53% of all sequences of length 20 generated by an unbiased random process. In contrast the occurrence rate for subsequence 0001 is around 0.75 meaning that this subsequence doesn't appear on only around 25% of unbiased sequences of length 20. Hahn & Warren (2009) argue that if human experience of unfolding random events mimics the sliding window, then this could explain three key tendencies of human randomness perception which are taken as evidence of bias:

i) a tendency to think that sequences with some irregularity are more likely given an unbiased coin
ii) an expectation of equal numbers of heads and tails within a sequence
iii) a tendency to over-alternate between outcomes when generating random sequences

Based on theoretical data of the kind presented here (figure 1B unbroken line), Hahn & Warren argue that i) is reasonable, i.e. the figure demonstrates that there is statistical support for the intuition that regular subsequences (e.g. 1111, 0101) occur less often than irregular subsequences (e.g. 0100, 1101). Hahn & Warren also argue that ii) is consistent with the sliding window account since it is difficult to distinguish between the vast majority of sequences using occurrence rate (figure 1B, unbroken line) suggesting judgments should be based not on an explicit coding of each subsequence but something simpler such as the proportion of heads. Finally Hahn & Warren argue iii) follows directly from the sliding window account since short sequences tend to have more alternations between outcomes than expected in an infinite series (Kareev, 1992).

Here we examine the characteristics of human random sequence generation in light of the theoretical account of Hahn & Warren (2009). To preempt our results we find that, in agreement with previous studies, human behavior departs markedly from that expected from a theoretical unbiased random generating process when compared on the *expected frequency of occurrence* of any binary sub-sequence. For an unbiased random process these expected frequencies should all be equal for any specified sub-sequence length. However, we also show that human sequences are remarkably similar to those of an unbiased random generation process when other methods of comparison are used which are relevant to the sliding window account (e.g. sub-sequence occurrence rate or direct comparison of sub-sequence frequency distributions for a given window length), and that this is particularly evident at sub-sequence lengths around 4 or 5 (i.e. a plausible length for a human window of experience as defined above).

Experiment

Participants first observed blocks of binary outcome random sequences following an unbiased Bernoulli process ($p = 0.5$) and were then instructed to generate random outputs to match the properties of the observed process.

Method

Participants. Twelve undergraduate students from the University of Manchester participated on a voluntary basis and gave informed consent. Participants received course credit as payment. There were no exclusion criteria.

Materials. Participants were seated in front of a 19-inch LCD display. The experimental stimuli were presented using the Python programming language on a PC running Windows 7. Participants responded using a standard Windows keyboard.

Design. We compared the statistical properties of sequences generated by a truly random Bernoulli process ($p = 0.5$) and those generated by our participants (N = 12) using four methods contrasting:

i) the expected frequency of sub-sequence occurrences per block of length 20
ii) the proportion of blocks of length 20 on which there was at least one sub-sequence occurrence. We call this the *occurrence rate* which is the complement of the *non-occurrence probability* described by Hahn & Warren (2009)
iii) occurrence frequency histograms for three sub-sequences of interest – perfect runs (e.g. 0000),

perfect alternations (e.g. 0101) and sequences such as (0001) which when compared to a perfect run of the same length has implications for the gambler's fallacy phenomenon

iv) boxplots illustrating medians and IQRs of occurrence frequency distributions for the three sub-sequences outlined in iii)

Procedure. Participants were told they would first observe the output of a machine generating a random sequence of 1's and 0's, and that they should attend to it (Presentation Phase) before going on to generate a sequence (Generation Phase).

Presentation Phase: Each digit (a 1 or 0) appeared on the screen for 250 msec before being replaced by the next digit in the sequence. The display of each digit was accompanied by a corresponding tone. The display was full screen with a black background. The digits were displayed in white in 80 point Arial font in the centre of the screen. 1's were accompanied by a 1200 hertz tone, and 0's by an 800 hertz tone. After every 20 digits the sequence paused and participants were required to complete a distractor task. The distractor task consisted of counting the number of vowels in a list of 10 words. In total participants observed 600 digits over 30 blocks of length 20.

Generation Phase: Participants were asked to generate a new sequence representative of the one they had just observed in the Presentation Phase. They used the keyboard to press either 1 with their left hand, or 0 with their right hand. For each key press participants saw the appropriate digit on screen and heard the corresponding tone, exactly as in the presentation phase. As in the Presentation Phase, participants generated 600 digits in 30 blocks of 20 and the same distractor task was used in between each block.

Analysis

We counted sub-sequences using sliding windows of lengths $k = 3$ to $k = 9$ and for global sequence length $n = 20$. For illustration we describe the analysis and present results for $k = 4$. For each participant, and each of the 30 blocks of data collected, we slid a window of length $k = 4$ through the $n = 20$ outcomes generated. We then undertook 4 analyses of these sequences:

Analysis 1 - Over the 360 (12 observers x 30 blocks) length 20 sequences, we calculated the expected value of the participant-generated frequency distribution for each of the 16 possible sub-sequences (0000, 0001,…,1111). For an unbiased random process the expected frequency of each sub-sequence should be 1.0625 per sequence of length 20.

Analysis 2 - Over the 360 (12 observers x 30 blocks) length 20 sequences, we calculated the occurrence rate – i.e. the proportion that contained at least one occurrence for each of the 16 possible sub-sequences (0000, 0001, …, 1111). Even for a random process this metric will not be the same for all sub-sequences since non-occurrence probabilities vary due to the sliding window analysis (see Hahn & Warren, 2009).

Analysis 3 - Over the 360 (12 observers x 30 blocks) length 20 sequences, we generated histograms illustrating the proportion of the 360 sequences containing 0, 1, 2, etc… occurrences of the three sub-sequences 0000, 0101, 0001.

Analysis 4 - Over the 360 (12 observers x 30 blocks) length 20 sequences, we generated boxplots illustrating the median and IQR of the distributions obtained in Analysis 3.

We generated the same amount of simulated data as that obtained from human participants using an unbiased Bernoulli process (p = 0.5). We refer to these simulated sequences as theoretical participant-generated and their properties are analyzed in an identical manner to the human data. By repeatedly generating (N = 1000) theoretical participant data sets we were able to place confidence bounds on the metrics described in Analysis 1 and 2 for the theoretical participant.

Results

In Figure 1A the dots represent the observed expected values of human-generated sub-sequence frequencies (Analysis 1) at window length 4. The unbroken black lines represent the equivalent metric for the theoretical participant. The dotted lines represent the 95% confidence interval on the theoretical data. Note that the theoretical expected frequencies are the same across sub-sequences since in an unbiased random process all sub-sequences at all window lengths should be equally represented (e.g. see Beltrami, 1999). Although the majority of the human data lies within the confidence interval for the theoretical participant, there are some clear departures and there appears to be systematic over and under-representation of certain sub-sequences relative to the theoretical participant. This analysis illustrates the standard description of human random sequence generation as biased. Relative to the theoretical participant, the perfect runs are clearly under-represented and 10 of the other 14 sub-sequences are over-represented.

Figure 1B shows the outcome of Analysis 2 for window length 4. The dots represent the occurrence rate – i.e. the proportion of the 360 blocks on which a sub-sequence occurred at least once – for human participants. Respectively, the solid black and dotted lines illustrate the equivalent occurrence rate and 95% confidence interval for the theoretical participant. Under this analysis the human and theoretical data share several common features, including a marked decrease in occurrence rate for perfect runs. In addition the human data appear to follow the fluctuations in the simulated data.

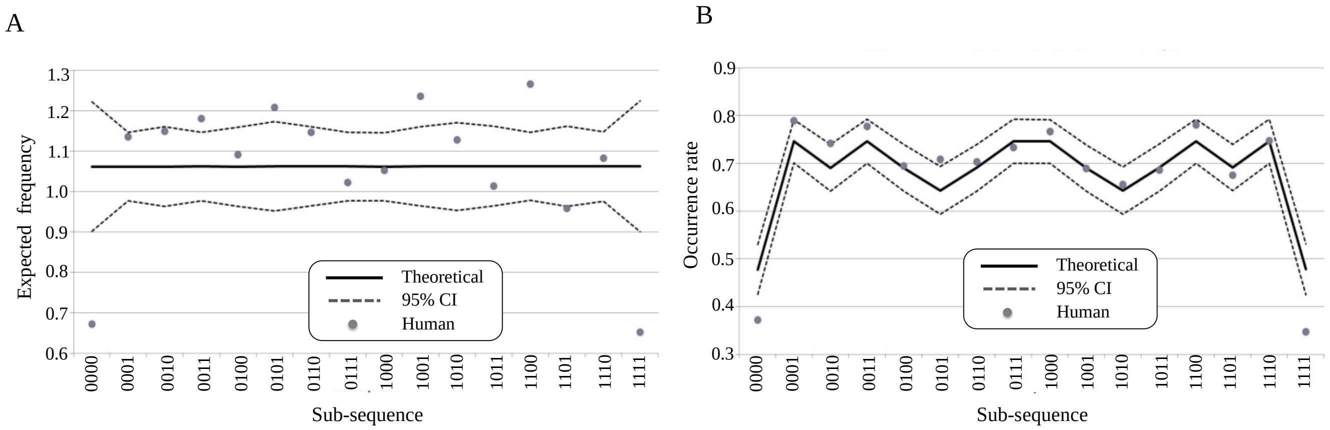

Figure 1: A. The results of Analysis 1 for sliding window length 4. Sub-sequence frequencies are presented for both human-generated (dots) and theoretical observer-generated length 20 sequences. B. The results of Analysis 2 for sliding window length 4. Proportions of blocks containing at least one occurrence of the sub-sequence are presented for both human-generated (dots) and theoretical observer-generated length 20 sequences

As noted in Hahn & Warren (2010), although the non-occurrence probability, or its complement the occurrence rate, is a convenient statistic with which to illustrate differences between sub-sequences it is not the only statistic for which differences emerge for an unbiased random process. In Analyses 3 and 4 we illustrate significant differences between the distributions, medians and modes of three key sub-sequences: 0000, 0001 and 0101 and show that based on these analyses human and theoretical data are in close agreement.

In figure 2 we present the outcome of Analysis 3 for the theoretical (figure 2A) and human (figure 2B) participants. Note, that occurrence rates obtained in Analysis 2 for the three sub-sequences considered can also be seen in figure 2 as the sum of all columns except that for frequency 0. Although there are some differences in the human vs. theoretical distributions they are primarily both qualitatively and quantitatively similar. Furthermore, the clear skew in the distributions of these data suggests that it is dangerous to use the expected value as a summary statistic (see Analysis 1). To further reinforce this point we have indicated the observed expected values (vertical dashed lines in figure 2). Note that for the theoretical data the expected values are identical at 1.0625 – which is consistent with Analysis 1. On the other hand for the human data the expected values are markedly different – again consistent with analysis 1. For example, note that the significant reduction in human relative to theoretical expected value for the sub-sequence 0000. However, this difference is largely driven by the fact that there are fewer high frequency sequences (e.g. beyond frequency 6) in the human data. These extreme values would contribute significantly to the expected value even though they are highly unlikely to be experienced. We suggest that placing emphasis on the difference in expected values between human and theoretical participants is problematic when there are similarities in the data generated on other (potentially more appropriate) statistics.

In figure 3 we present another illustration of the data in figure 2. These boxplots emphasize the similarity in the median frequency for the humans and theoretical data. In addition, box plots for the 0001 and 0101 sub-sequences are very similar between human and theoretical participants. Similar to figure 2, for sub-sequence 0000 the increased tendency for the theoretical participant to generate high frequency sequences is also evident. As noted above, this tendency is responsible for the higher expected value for theoretical relative to human data. In addition we see that for an agent paying attention to the median statistic it would be true to say that sub-sequence 0001 is less likely to occur that 0000. It is possible that this plays a role in the gambler's fallacy.

Note that although we have focused exclusively on the analyses at window length k = 4 we have data for lengths from k = 3-9. We find that up to length 4 or 5 there is good correspondence between human and simulated data on Analyses 2, 3 & 4 but beyond this value the discrepancies are greatly increased.

Discussion

The purpose of the present study was to provide a preliminary test of the theoretical account of randomness perception put forward by Hahn & Warren (2009). In particular we wanted to go beyond the standard account which presents a picture of randomness perception as highly biased because the frequencies of human-generated sub-sequences depart from those expected from a truly random process (figure 1A). Instead, we present a set of alternative analyses under which human performance is comparable to that of a random process.

The key result here is that the correspondence between human and unbiased theoretical data depends on the statistics used to parameterize performance. We have presented several analyses that emphasize the similarities.

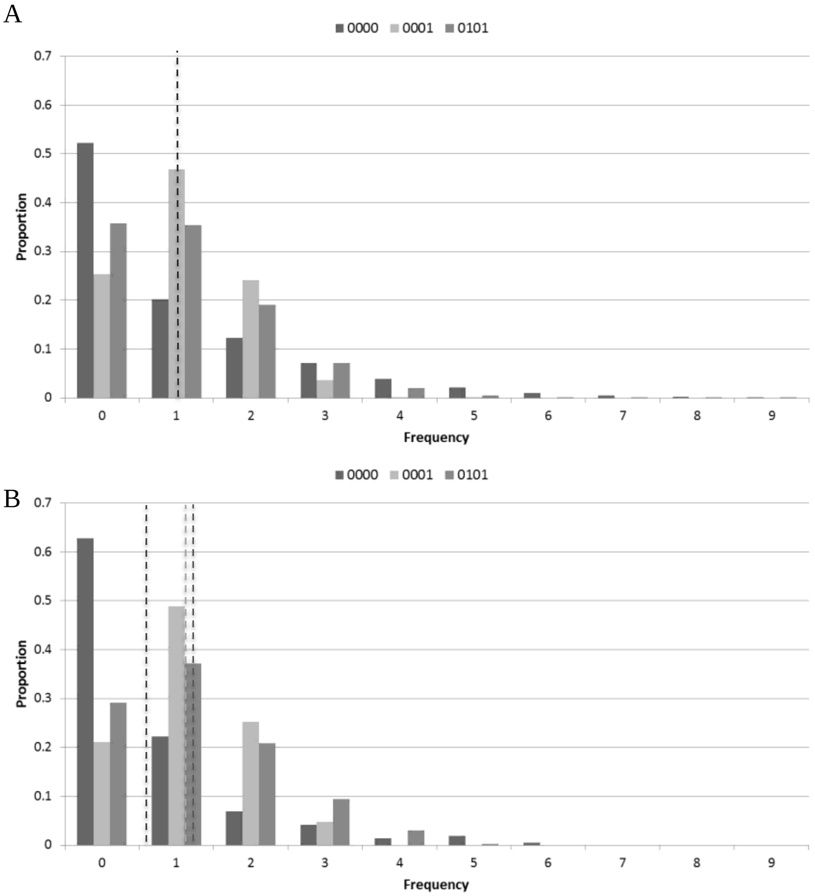

Figure 2: The results of Analysis 3 for sliding window length 4. Histograms describing proportion of blocks containing each frequency for three selected sub-sequences. Vertical dashed lines represent the expected values of the distributions. A. Data for theoretical participant (truncated at frequency 9). Note the expected values overlap (consistent with Analysis 1). B. Data for human observers.

Moreover, the analyses conducted are sensible in that they reflect the manner in which we are likely to experience random events due to the constraints imposed on human cognition – i.e. as a sliding window moving one outcome at a time through a longer but finite sequence of unfolding events.

The results presented re-emphasize the argument made in Hahn & Warren (2010) that the mean (expected value) is not an appropriate statistic to characterize the distribution of sub-sequences generated by either a human or theoretical participant under a sliding window analysis. The level of skew in the data is high and it is precisely for such distributions that the median and/or mode are preferable. As noted in Hahn & Warren (2010), it would seem problematic to conclude that average income was $100,000 per month in a population where most made $1000 and very few made $1,000,000. By the same logic, based on the distribution presented in figure 2, it is not sensible to suggest that one would expect to see (on average) about one instance of HHHH in 20 coin flips. In contrast the median (figure 3) and or/mode (figure 2) statistics are more meaningful, and, based on these statistics humans look rather well matched to the unbiased theoretical process.

The fact that human and theoretical sequence generation processes share common features for Analyses 2-4 at window lengths 3-5 suggests that it is possible that on average our participants were behaving similarly to the process described in Hahn & Warren (2009) with sliding window length around 4-5. In practice, individuals are likely to have different and possibly non-stationary sliding window lengths. If enough data is generated, it may be possible to establish a link between individual sequence statistics and a proxy measure of window length such as digit-span or short-term memory capacity. An investigation of this possibility will form the basis of future work.

In summary we provide experimental data that is consistent with the account put forward by Hahn & Warren (2009; 2010). We suggest that apparent biases in human

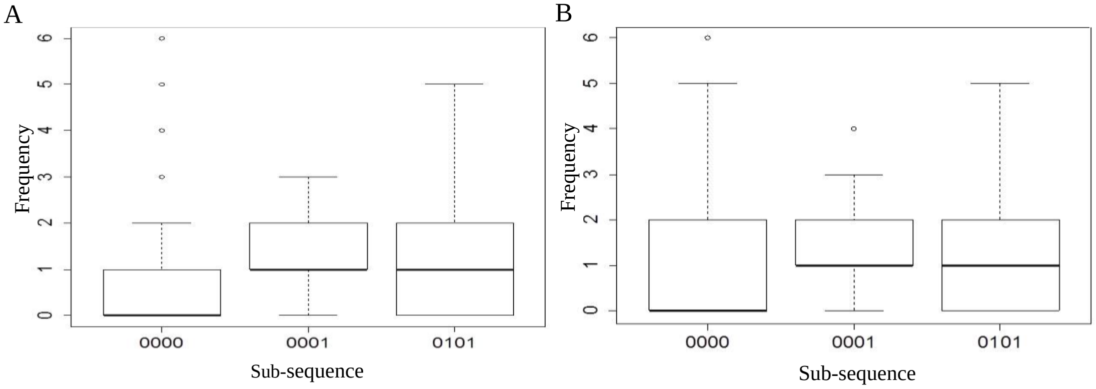

Figure 3: The results of Analysis 4 for sliding window length 4. Boxplots illustrating medians IQRs and extreme values of the data illustrated in figure 2 for three selected sequences. A. Data for theoretical participant (truncated at frequency 6). B. Data for human observers.

randomness perception should be re-evaluated and that it is problematic to suggest human behaviour is flawed simply because it departs from that of an unbiased theoretical process on a single metric which may not reflect cognitive and task constraints.

Acknowledgments

This research was funded by ESRC transformative research grant ES/L00321X/1 awarded to PAW, UH, AH, WED.

References

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

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

Beltrami, E. (1999). *What is random? Chance and order in mathematics and life*. New York: Springer.

Clotfelter, C.T & Cook, P.J. (1993). The "Gambler's Fallacy" in Lottery Play. *Management Science, 39*, 1521-1525.

Croson, R. & Sundali, J. (2005). The Gambler's Fallacy and the Hot Hand: Empirical Data from Casinos. *Journal of Risk and Uncertainty, 30,* 195-209.

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology , 62,* 385-394.

Hahn, U. (2011) *The Gambler's Fallacy*, Oxford Bibliographies Online.

Hahn, U. & Warren P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review, 116, 454-461.*

Hahn, U. and Warren, P. A. (2010). Why three heads are a better bet than four: a reply to Sun, Tweney, and Wang (2010). *Psychological Review, 117, 706-711.*

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kareev, Y. (1992). Not that bad after all: Generation of random sequences. *Journal of Experimental Psychology: Human Perception and Performance , 18*, 1189-1194.

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

Lopes, L.L. & Oden, G.C. (1987). Distinguishing Between Random and Nonrandom Events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 392-400.

Nickerson, R.S. (2002). The Production and Perception of Randomness . *Psychological Review , 109*, 330-357.

Nickerson, R.S. & Butler, S.F. (2009) On producing random binary sequences. *The American Journal of Psychology, 122*, 141-151.

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

Oskarsson, A.T., van Boven, L. Oskarsson AT, Van Boven L, McClelland GH, Hastie R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin, 135*, 262-85.

Rapoport, A. & Budescu, D. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General, 121*, 352-363.

Terrell, D. (1998). Biases in Assessments of Probabilities: New Evidence from Greyhound Races. *Journal of Risk and Uncertainty, 17,* 151–166.

Toneatto, T., Blitz-Miller, T., Calderwood, K., Dragonetti, R. & Tsanos, A. (1997). Cognitive distortions in heavy gambling. *Journal of Gambling Studies, 13*, 253-266.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of the literature. *Psychological Bulletin, 77,* 65–72.

Williams, J. J., & Griffiths, T. L. (2013). Why are people bad at detecting randomness? A statistical argument. *Journal of Experimental Psychology: Learning, Memory & Cognition, 39*, 1473-1490.