Why are people bad at detecting randomness? A statistical analysis

Joseph J. Williams and Thomas L. Griffiths
Department of Psychology
University of California, Berkeley

Word count: 11 681

Address for correspondence:
Joseph Jay Williams
University of California, Berkeley
Department of Psychology
3210 Tolman Hall # 1650
Berkeley CA 94720-1650
E-mail: joseph_williams@berkeley.edu
Phone: (510) 501 7884 **Fax:** (510) 642 5293

Errors in detecting randomness are often understood in terms of biases and misconceptions, but we propose and provide evidence for a *nested-process* account that characterizes the contribution of the inherent statistical difficulty of the problem. The account uses a Bayesian statistical analysis to reveal that a random process is contained or nested within a range of systematic processes. The consequence is that randomly generated data are still reasonably likely to have come from a systematic process, and thus only weakly diagnostic of a random process and easy to misidentify. Experiments 1 and 2 show that the low accuracy in judging whether a sequence of coin flips is random (or biased towards heads or tails) is due to the weak evidence provided by random sequences. While randomness judgments were less accurate than non-nested judgments in the same task domain, once the strength of the evidence available was statistically equated, accuracy on these non-nested judgments was reduced and was no longer significantly different. Experiment 3 extended this finding to assessing whether a sequence was random or exhibited sequential dependence, and found that the nested feature of randomness caused errors in addition to those stemming from known misconceptions. Overall, judgment accuracy was better predicted by whether or not a judgment involved a nested process than whether or not it concerned a random process.

Does the admission of four men and one woman to a graduate program reflect gender discrimination, or just random variation? Are you more likely to give a good presentation if your last presentation went well, or are they independent of each other? Do people who take vitamins get sick any less often than people who do not? People have a remarkable capacity to detect patterns, and both children and adults are often preoccupied with finding the statistical regularities, causal structures, and predictive relationships that exist in the world. The alternative to finding structure in the environment is to realize that it is unsystematic: Certain events and patterns occur in the absence of systematic forces and the processes that generate them are random. Discriminating between events that occur at random and observations that provide evidence for underlying structure is important in learning about the actual structure of the world.

Despite the value of this capacity to detect statistical relationships, a great deal of psychological research has pointed out the errors that people make in discerning the presence of random rather than systematic processes. People detect streaks in sequences of events which exhibit no sequential dependence (Gilovich, Vallone, & Tversky, 1985), and the pervasive phenomenon of *illusory correlation* reflects that people who observe randomly co-occurring variables may incorrectly infer correlations or causal relationships (Chapman & Chapman, 1967, Hamilton, 1981, Redelmeier & Tversky, 1996). A broad range of studies have investigated the flaws

This work was supported by grants FA9550-07-1-0351 and FA-9550-10-1-0232 from the Air Force Office of Scientific Research to TLG, and a post-graduate scholarship from the National Science and Engineering Research Council of Canada to JJW.

that underlie the human understanding of randomness and the reluctance to recognize its existence, providing a compelling account of why people are so poor at detecting randomness and generating observations randomly. Although misconceptions and biases are clearly an important factor in people's erroneous judgments about randomness, a further challenge they face may be in the mathematical nature of the problem. It is notoriously difficult to define or prove the presence of a random process, and previous work has emphasized that judging randomness may be inherently difficult (Lopes & Oden, 1987, Nickerson, 2002), and that evaluating the judgment abilities of naive and biased reasoners should proceed by comparison to normative standards (Lopes, 1982). However, little research has directly addressed this question through computational modeling and experimental work (but see Lopes, 1982).

In this paper, we aim to formally characterize the statistical challenges in detecting randomness – which pose a formidable problem for both people and ideal reasoners free of biases and processing constraints. The paper presents a *nested-process* account, according to which detecting a random process is hard because it is *nested* or contained as a limiting case in the range of systematic processes. We propose that this imposes statistical limitations on how diagnostic randomly generated data can be, as it can always be accounted for by a systematic process. Before presenting our formalization and empirical tests of the account, we briefly review past work on people's difficulties in evaluating randomness.

Errors in generating and judging randomness

Early research that solicited people's production of random binary sequences (such as heads and tails) revealed that people produce sequences that contain excessively balanced numbers of heads and tails and too few repetitions (for a review see Wagenaar, 1972). Furthermore, many sequences that are in fact randomly generated are judged as reflecting a systematic bias towards heads or tails, or towards repetition. Errors in evaluating randomness manifest themselves in real world settings: The *gambler's fallacy* refers to the mistaken belief that any systematicity in a randomly generated sequence will be "corrected" (Tune, 1964; Kahneman & Tversky, 1972), and roulette gamblers express the belief that a red result becomes more likely after a run of black.

This pattern of errors can be summarized in terms of an *alternation bias*, whereby people perceive alternations (changing from heads to tails or vice versa) as more indicative of randomness than repetitions (repeating a head or tail), so that sequences with an alternation rate of 0.6 or 0.7 are incorrectly perceived as "most random" (Falk & Konold, 1997; Lopes & Oden, 1987; Rapoport & Budescu, 1992). This bias even influences what is remembered about random sequences (Olivola & Oppenheimer, 2008). A related topic of debate (Gilovich et al., 1985, Alter & Oppenheimer, 2006) has been whether belief in the *hot hand* effect reflects people's tendency to incorrectly detect sequential dependencies in events that are actually sequentially independent. Nickerson (2002), Falk and Konold (1997), and Bar-Hillel and Wagenaar (1993) provide reviews of research on randomness, giving further details on the errors people make.

Why do people make so many errors in judgments about randomness? Biases such as those reviewed above suggest the possibility that people's intuitions about randomness are flawed: Bar-Hillel and Wagenaar (1993) suggest that "People either acquire an erroneous concept of randomness, or fail to unlearn it." (p. 388). One influential account of the nature of the error is based on the idea that people evaluate randomness by judging how *representative* some observations are of a random process (Kahneman & Tversky, 1972). Instead of calculating the probability of observations under a random process, people rely on the heuristic of judging to what extent these observations represent the essential characteristics of random data. The concept is further elaborated to that of *local representativeness*: People expect even small samples to closely represent the properties of randomly generated data, although small randomly generated samples often contain structure by chance. Other accounts include limitations on people's memory capacities (Kareev, 1992, 1995), construal of randomness in terms of ease of processing (Falk & Konold, 1997), and the suggestion that the use of ambiguous or misleading instructions may underestimate people's abilities (Nickerson, 2002).

The statistical challenge underlying randomness judgment

In the remainder of the paper we consider the abstract statistical problem involved in identifying whether an observation is random, and use this analysis as the basis for a *nested-process* account of the inherent difficulty of detecting randomness. We motivate our account's key insight about *nested* processes in the context of an example, then present the framework for our ideal observer analysis of the computational problem of randomness judgment. We then explain and derive predictions from our model of discriminating nested processes, and carry out three experiments that test the model predictions about how nested processes provide only weakly diagnostic observations and lead to errors in judgment.

The paper will focus on tasks with the formal structure of deciding whether two outcomes are random in the sense of being equally likely to occur or systematic in that one is more likely than the other. A wide range of real-world judgments in different domains and contexts have this abstract form. We will discuss two such judgments. The first concerns the relative frequency of two events. For example, determining whether men and women are equally likely to be admitted to a graduate program, whether two students perform equally well on exams or one does better than the other, and whether a coin is fair or biased to heads or tails. The second concerns sequential dependence between successive events. When there are two equally likely events, the outcome of interest is then whether an occurrence of an event is followed by a repetition of the event or an alternation to the other event. Judging randomness therefore involves assessing if events are random in being sequentially independent (the outcomes of repetition and alternation are equally likely) or sequentially dependent (one outcom – e.g., repetition – is more likely than the other). For example, if there is no gender bias in graduate admission, is there a relationship between the gender of successive admittees? For a fair coin, are the flips are random (independent) or does a head (tail) on one flip influence the next?

Consider the first scenario, examining ten candidates to evaluate whether admissions are gender neutral – random with respect to being male or female. Judgment accuracy could be reduced by misconceptions about randomness or the use of biased heuristics. But there is also a subtle but significant statistical challenge in this problem, which we predict will cause judgment errors even in the absence of misconceptions and with even unlimited processing resources. If the gender distribution is random then $P(\text{male})$ is 0.5, while if it is systematic $P(\text{male})$ is somewhere in the range from 0 to 1. If six males and four females are admitted, this might seem to provide evidence for a random process. But how strong is the evidence? In fact, six males and four females could also be produced by a systematically biased process, one in which $P(\text{male})$ is 0.6, or even 0.55 or 0.7. While likely under a random process, the observation can *also* be explained by a systematic process, and so it is only weakly diagnostic of a random process and leads to inaccuracy. The problem is that a random process is *nested* as a special case within the broad range of systematic processes.

Formalizing the inference problem

To formally investigate the nature of the statistical challenge present in randomness judgment, an ideal observer model or rational analysis (in the spirit of Anderson, 1990) can be used characterize to the problem a reasoner faces in evaluating randomness. The formal framework can be applied to a range of contexts, but for the purposes of this paper we consider evaluating whether some data set d of binary outcomes is random (equiprobable) or systematic (not equiprobable). We discuss it in the context of evaluating whether sequences of coin flips are random or not, a task that affords experimental control and has been extensively investigated in previous literature. The model addresses two aspects of randomness: (1) evaluating whether a coin is random in being equally likely to give heads or tails, vs. weighted towards heads over tails (or vice versa), and (2) even if heads and tails are equally likely, evaluating whether a coin is random in being equally likely to repeat or alternate flips (sequential independence), vs. more likely to repeat or to alternate (sequentially dependent).

The hypotheses under consideration are represented as:

h_0: The data were generated by a random process. For example, $P(\text{heads}) = 0.5$ or $P(\text{repetition}) = 0.5$.

h_1: The data were generated by a systematic process. For example $P(\text{heads})$ (or $P(\text{repetition})$) follows a uniform distribution between 0 and 1.[1]

A rational solution to the problem of evaluating these hypotheses in light of data is to use Bayesian inference. In this case, we can write Bayes' rule in its "log odds" form:

$$\log \frac{P(h_1|d)}{P(h_0|d)} = \log \frac{P(d|h_1)}{P(d|h_0)} + \log \frac{P(h_1)}{P(h_0)}. \qquad (1)$$

This equation says that the relative probability of a random (h_0) or systematic (h_1) process after seeing data d (denoted by $\log \frac{P(h_1|d)}{P(h_0|d)}$) depends on how likely the data d are under a random process versus a systematic process ($\log \frac{P(d|h_1)}{P(d|h_0)}$), and how likely either process was before seeing the data ($\log \frac{P(h_1)}{P(h_0)}$). For the purposes of this paper, the key term in Equation 1 is the *log likelihood ratio* $\log \frac{P(d|h_1)}{P(d|h_0)}$, which quantifies the strength of evidence the data provide for h_1 versus h_0.[2]

To calculate these terms and report our model simulation, we consider a case where the observed data consist of ten outcomes (10 head/tail coin flips or 10 repetitions/alternations). The number of heads (repetitions) in each batch of 10 follows a Binomial distribution. For sequences from a random process, the probability of a head (repetition) is 0.5. For systematic processes, it ranges uniformly from 0 to 1. This task is sufficiently specified to compute the log-likelihood ratio introduced in Equation 1, which provides a quantitative measure of the evidence a data set provides for a random vs. systematic process (see Appendix A for details). We now illustrate the insights of the model in the context of evaluating whether sequences reflect a coin for which heads and tails are equally likely (vs. weighted to one over the other), although the results also apply to the mathematically equivalent task of evaluating sequential independence in repetitions and alternations.

Randomly generated data sets provide only weak evidence for randomness

The key results of our ideal observer analysis are presented in Figure 1. Figure 1 (a) shows how likely different data sets of ten flips are to be generated by each process, as a function of the number of heads in the data set. The horizontal axis gives the number of heads and tails in a data set of ten flips. The vertical axis gives the probability of a data set being generated, where the black line represents $P(d|h_0)$ (the probability the data set would be generated from a fair/random coin) and the grey line $P(d|h_1)$ (the probability the data set would be generated from a systematically biased

[1] Intuitively, it might seem the hypothesis of systematicity should exclude the hypothesis of randomness (e.g. $P(\text{heads})$ between 0 and 1 except for 0.5). Representing the hypothesis of systematicity in this way is mathematically equivalent to the current formulation, and makes no difference to any of our conclusions.

[2] This log likelihood ratio has been used in other mathematical definitions of randomness (Griffiths & Tenenbaum, 2001), and has also been proposed as a measure of the representativeness of an observation relative to a hypothesis (Tenenbaum & Griffiths, 2001).

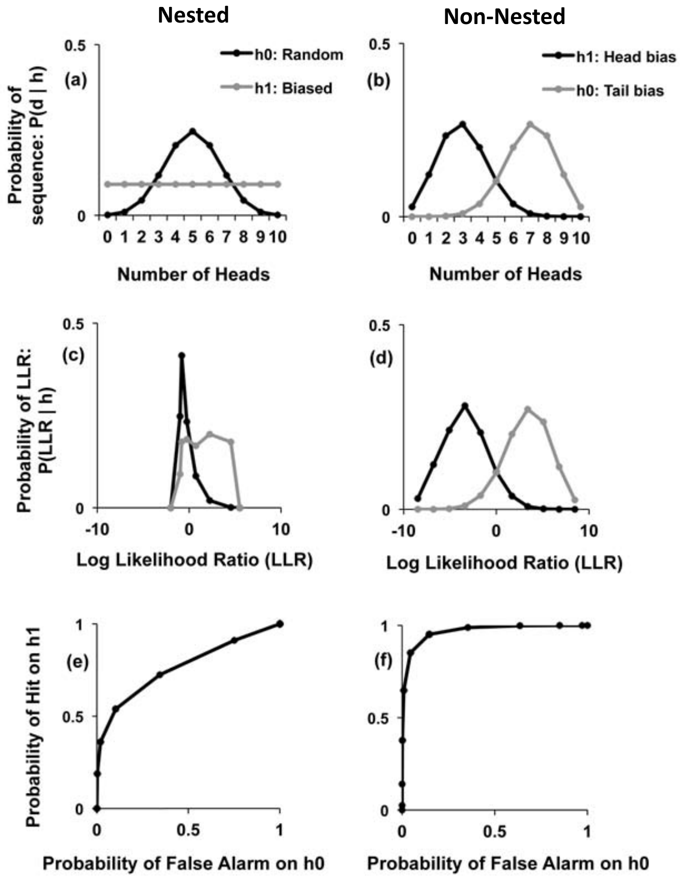


Figure 1. Left column shows the model of the nested problem of randomness judgment: discriminating whether events like head/tail flips or repetition/alternation of flips are equiprobable or systematically biased. Right column shows the model of the non-nested problem of discriminating the direction of systematic bias. Plots show the probability distribution over sequences for nested (a) and non-nested (b) processes, the distribution of evidence (measured by the log likelihood ratio or LLR) for each of the nested (c) and non-nested (d) processes, and the ROC curves for the nested (e) and non-nested (f) discrimination tasks.

coin). Data sets with little or no systematic bias are likely to come from a random process (e.g., 5H5T, 6H4T), while data sets with a wide range of systematic bias are likely under a systematic process (e.g., 0H10T to 10H0T).[3] However, all of the data sets likely to be generated by a random process are also reasonably likely to come from a systematic process, while the converse is true for only some systematically generated data sets (e.g., a random process is very unlikely to generate a sequence with 9H1T). This is because a random process is a special case of a systematic process (a $P(\text{heads})$ of 0.5 is a point in the range 0 to 1): A random process is contained – more formally, *nested* – in the set of systematic processes.

Recall that the log likelihood ratio (LLR) ($\log \frac{P(d|h_1)}{P(d|h_0)}$) serves as a quantitative measure of the relative evidence a data set d provides for a systematic versus random process. It quantifies the relative probability of the sequence being generated by one process ($P(d|h_1)$ for systematic) rather than the other ($P(d|h_0)$ for random). Figure 1 (c) shows the distribution of evidence (the distribution of the LLR) for sequences generated from the random process and sequences from the systematic process.

We explain the construction of these distributions to aid in their interpretation. The distribution of the LLR for a random process was obtained as follows. First, 5000 sequences of 10 coin flips were generated from the distribution associated with h_0. For each of the 5000 sequences the LLR was calculated, and these 5000 LLRs were used to create the relative frequency plot in Figure 1 (c) (h_0: black line). The details of calculating the LLR in this case are given in Appendix A. In Figure 1 (c), the horizontal axis displays the range of LLRs different sequences can have (calculated with respect to the hypotheses about a random versus systematic process). The vertical axis depicts how likely sequences with these LLRs are. An analogous procedure was used to construct the distribution of the LLR for h_1: 5000 sequences were generated from a systematic process (for each sequence, $P(\text{heads})$ was randomly sampled from a uniform distribution between 0 and 1), and the LLRs of all 5000 sequences were calculated and used to create a relative frequency plot (h_1: grey line).

Figure 1 (c) shows that the majority of randomly generated sequences have small negative LLRs (e.g., the LLR of 5H5T is -1.0). While a negative LLR indicates that the sequence is more likely to be generated by a random than systematic process, the size or magnitude of the LLR indicates *how much* more likely this is. The greater the magnitude of the LLR for a sequence, the stronger the evidence the sequence provides for one process over the other. Sequences with LLRs near to zero provide little evidence as either process is likely to generate them. While there are some systematically generated sequences with small LLRs, there are many that have large positive LLRs (e.g. the LLR of 10H0T is 4.5) and so provide strong evidence for a systematic process. Throughout this paper the LLR provides a precise quantitative measure of the evidence a data set provides for one process versus another. The results validate the nested-process account: one consequence of a random process being nested in a range of systematic processes is that randomly generated data can provide only weak evidence for a random process.

Comparison to non-nested processes

Throughout this paper, we spell out the distinctive challenges of judgments of nested processes (and by extension randomness judgments) by comparing them to judgments about non-nested processes. We examine non-nested processes whose probability distributions over data sets have a

[3] Note that because such a broad range of data sets are likely under a systematic process, a lower probability must be assigned to each of them.

similar shape and only partially overlap. Such non-nested processes appear more frequently than nested processes in psychology, and are often assumed in signal detection tasks like deciding whether an item on a memory test is old or new or identifying a perceptual stimulus in a noisy environment (Green & Swets, 1966).

One judgment about binary outcomes that involves non-nested processes concerns which outcome the generating process is biased towards. A simple version of this might compare the hypotheses that a coin is biased towards heads (h_0: $P(\text{heads}) = 0.3$) versus biased towards tails (h_1: $P(\text{heads}) = 0.7$).[4] Figure 1 (b) shows how likely different sequences of 10 coin flips are under these two processes. Again, the horizontal axis depicts particular sequences (e.g., 2H8T, 4H6T) and the vertical axis gives the probability of the sequence being generated by a process biased towards tails (h_0: black line) and a process biased towards heads (h_1: grey line). A comparison of the nested processes in Figure 1 (a) and the non-nested processes in Figure 1 (b) reveals key differences. While sequences that are likely under both non-nested processes (e.g., 5H5T, 6H4T) are ambiguous, neither process is nested within the other, and so each process can generate sequences that are very unlikely to come from the other process.

The distribution of the LLR for sequences generated from non-nested processes is shown in Figure 1 (d), and was constructed using a similar procedure to Figure 1 (c). First, 5000 sequences were generated from a process biased towards tails ($P(\text{heads}) = 30\%$) and 5000 from a process biased towards heads ($P(\text{heads}) = 70\%$). The LLR of each sequence was computed as $\log \frac{P(d|h_1)}{P(d|h_0)}$. It should be noted that these are not the same probabilities used for the nested processes, because h_0 now represents a bias towards tails instead of a fair coin ($P(\text{heads}) = 0.3$, not 0.5), and h_1 a bias towards heads instead of any bias ($P(\text{heads}) = 0.7$, not a uniform distribution from 0 and 1). The formula for the LLR is provided in Appendix A. The relative frequency plot in Figure 1 (d) shows the distribution of the sequence LLRs, where the horizontal axis depicts the LLRs of particular sequences (calculated with respect to the hypotheses of a tail bias versus head bias) and the vertical axis depicts how likely sequences with these LLRs are.

Although the LLR of a sequence is calculated with respect to different hypotheses in Figures 1 (c) and 1 (d), the LLR still permits a direct comparison of the strength of the available evidence. The LLR's value is to serve as an abstract and context-independent quantification of the evidence a data set provides in discriminating any two given hypotheses. For example, the sequence 5H5T has an LLR of -1.0 with respect to whether the generating coin was fair or biased (and an LLR of 0 with respect to whether it was biased to tails or heads) while the sequence 4H6T has an LLR of -1.7 with respect to whether the generating coin was biased to tails or heads (and an LLR of -0.8 with respect to whether the coin is fair or biased).

A comparison of Figure 1 (c) and (d) demonstrates the distinctive statistical challenge that stems from the nested nature of randomness judgment. Whereas the distribution of evidence for nested processes is asymmetric and substantially weaker for the nested random process, the distribution of evidence for non-nested processes is symmetric and a broad range of sequences provide strong evidence for the process that generated them. Randomness judgment not only differs from many standard judgment tasks in requiring people to draw on concepts of and reasoning about a random process, but *also* in being a nested judgment and therefore having statistical limits on the evidence available.

Difficulty of discrimination as reflected in ROC curves

To quantify judgment accuracy for nested and non-nested processes we draw on tools from signal detection theory. Signal detection theory is useful in quantifying the difficulty of judgment tasks across a range of situations, and Lopes (1982) argues for its value in understanding randomness judgment – particularly in comparing human reasoners to a normative standard. We examine the *receiver operating characteristic* or *ROC* curve for nested and non-nested judgments. To infer from a sequence whether h_0 or h_1 is true a reasoner must have a decision criterion based on the evidence – for example they could report h_0 whenever the LLR is below zero and h_1 when it is above. However, the criteria adopted can vary across prior expectations of the likelihood of h_0 and h_1, different costs and rewards for errors and correct responses, and individuals. We use the ROC curve because it provides a broad view of how difficult or easy it is to use a sequence to discriminate two processes, without relying on a specific judgment criterion. The ROC curve for discriminating the nested random and systematic processes is shown in Figure 1 (e), and the ROC curve for discriminating the two non-nested systematic processes in Figure 1 (f).

The details of how these curves were constructed are provided in Appendix B, but the curve in Figure 1 (e) plots the relative proportion of *hits* (correct identifications of systematically generated data sets) on the vertical axis against the proportion of *false alarms* (misclassification of randomly generated data sets as systematic). If only a single criterion were used (e.g. an LLR of 0) this curve would collapse to a single point that plots the predicted hit rate against the false alarm rate. However, we calculated the hit rate and false alarm rate for many criteria that cover a broad range (from conservative to liberal in reporting h_1) to produce these curves. Each ROC curve therefore gives a broad and criterion-independent picture of an ideal observer's ability to use the evidence available to discern which process generated a sequence. Curves which are closer to a right

[4] The conclusions of this analysis are not significantly changed by manipulating these particular probabilities (e.g., using $P(\text{heads})$ of 0.25 or 0.35) as long as it represents a reasonable bias. For example, $P(\text{heads}) = 0.52$ is a less plausible representation of participants' belief that a coin is biased towards heads than $P(\text{heads}) = 0.70$. Representing bias over a uniform interval (e.g., $P(\text{heads})$ ranges uniformly from 0.5 to 1) also produces equivalent results, as is later demonstrated in the model in Figure 2.

angle demonstrate good discriminability of the two processes, while curves that are closer to the diagonal reflect reduced ability: Increasing hits requires large increases in false alarms.

Even if misconceptions about random processes are absent and cognitive resources are not taxed, the ROC curves show that discrimination accuracy is inherently lower for the nested than the non-nested processes, which is caused by the weaker distribution of evidence. The ROC curves also emphasize that randomness judgment involves an inherent tradeoff between accurately identifying random processes and accurately identifying systematic processes – increasing detection of systematicity *necessitates* mistakenly claiming that randomly generated data reflect a systematic process. Since the weak evidence reduces discriminability, reasoners will be especially prone to erroneously detecting structure when the data is randomly generated – the key phenomenon identified in past research.

Summary

The ideal observer analysis elucidates the precise nature of the inherent statistical difficulty in detecting randomness – it is a nested process. Discriminating a random process (like $P(\text{heads})$ or $P(\text{repetition})$ is 0.5) from a systematic process ($P(\text{heads})$ or $P(\text{repetition})$ has another value between 0 and 1) is statistically difficult because, as demonstrated in the analysis of the likelihood of data sets, randomly generated data sets are also reasonably likely to have come from systematic processes. Calculating the distribution of the LLRs of data sets generated by both kinds of processes provided a quantitative measure of the evidence a data set provides, demonstrating that randomly generated data sets provide relatively weak evidence for a random process. The paucity of this evidence was outlined in the comparison to the evidence that can be provided for a systematic process, and to the evidence provided by data sets from non-nested processes. ROC curves indicated that the information available in judging randomness was lower than for the other tasks, such that raising correct identifications of systematic process would necessitate higher false alarms in incorrectly judging that randomly generated data set reflected a systematic process.

One concern with the current model may that the non-nested processes are rendered easier to discriminate by selective choice of the parameter values of $P(\text{heads})$ or $P(\text{repetition})$ of 0.3 and 0.7. To address this concern we confirmed that the challenge in nested judgments was also apparent when compared to another model for non-nested processes. In this model $P(\text{heads})$ ($P(\text{repetition})$) ranged from 0 to 0.5 for h_0 and 0.5 to 1 for h_1. This used the same assumptions as the previous simulation, and the model's derivation is shown in Appendix A. As Figure 2 shows, changing these parameters may adjust the distribution of the LLR (evidence), but because neither of these processes is nested in the other, the unique challenge faced in judging randomness still remains. We now present three experiments that test whether this model does in fact predict people's errors in detecting randomness.

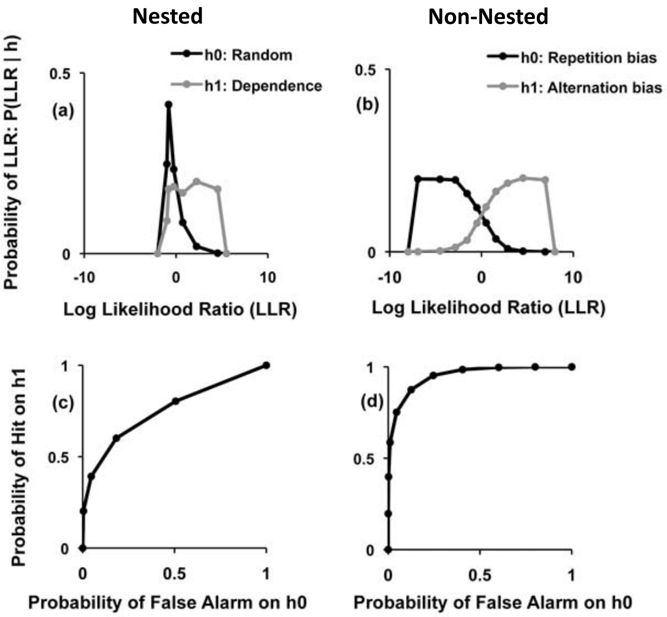


Figure 2. Left column replicates the nested model from Figure 1. Right column shows the non-nested model where systematic processes are distributed over intervals 0 to 0.5 and 0.5 to 1. Distributions of the LLR for these nested (a) and non-nested (b) processes, and ROC curves for discriminating nested (c) and non-nested (d) processes.

Exploring the source of errors in human randomness judgments

Our nested-process account provides a novel proposal for how to characterize the statistical difficulty of making randomness judgments. But there is no empirical evidence that people's judgments are actually sensitive to the statistical measures we present. Moreover, there is clear reason to believe people have misconceptions about randomness and processing limitations, which may eliminate or overwhelm any effects of our statistical measures on judgment.

We conducted three experiments that investigated the extent to which accuracy and errors depended on statistical features characterized by the model – like the LLR or quantity of evidence available – versus whether people needed to reason about and represent a random process. Our model predicts that accuracy should be primarily a function of the evidence provided by a sequence (LLR), which our mathematical analysis shows is highly dependent on whether a process is nested or non-nested. Alternatively, the model may fail to accurately capture the evidence available to people, or statistical considerations may play a minimal role if errors are driven largely by people's difficulties in conceptualizing and reasoning about randomness.

All three experiments compared the accuracy of judgments in a *nested* condition – discriminating a random from a systematically biased process – to judgments in a *non-nested* condition – discriminating two systematic processes. Accuracy is predicted to be lower in the nested condition, whether because of (1) people's limitations in conceptualizing and reasoning about a random process; and/or (2) the low LLRs or weak evidence available for a nested process – as predicted by our model. To evaluate these possibilities we compared the nested and non-nested condition to a critical *matched* condition. The matched condition used the same judgment task as the non-nested condition, but the same distribution of evidence as the nested condition. Although people did not need to reason about a random process, the model were used to statistically equate the available evidence to that in the nested condition. The model's predictions about the LLR was used to choose the sequences in the matched condition so they provided exactly as much evidence for discriminating the non-nested processes as the sequences in the nested condition did for discriminating random from systematic processes.

If our nested-process account characterizes the statistical difficulty people face in detecting randomness, the matched condition should have lower accuracy than the non-nested condition. If the model captures the difficulty of the task, the matched condition may even be as inaccurate as the nested. If the model does not capture difficulty or these considerations are minimal relevant to other factors, accuracy in the matched condition should not differ from the non-nested and could even be greater. The model can also be evaluated by assessing how well the LLR – the model's measure of evidence – predicts people's accuracy and reaction time in making judgments on particular data sets. The model predicts that judgments on sequences with small LLRs (not very diagnostic) should be near chance and have slow reaction times, with the opposite pattern for sequences with large LLRs.

While all the experiments followed this basic logic, the task in Experiments 1 and 2 was deciding if a coin was random (heads and tails equally likely) or biased towards heads/tails. Experiments 3A and 3B extended the model to the more complex context of deciding whether a coin was random (independent of previous coin flips – repetitions or alternations equally likely) or biased towards repetition/alternation. These experiments also extended the current model to investigate whether the nested-process account predicts people's judgment errors even in situations in which there are known biases and misconceptions that cause errors. The model could then shed light on how to integrate both rational considerations and erroneous ideas about random processes in fully explaining why randomness judgment is so hard. Alternatively, invoking a further statistical account might be unnecessary and lack parsimony, or the strength of the evidence may play a minimal role relative to flawed reasoning about randomness.

Experiment 2 also went beyond the other two experiments by independently manipulating whether judgments did or did not concern a random process, and whether they were nested or non-nested. This aimed to further elucidate whether errors were best predicted by whether people have to reason about random processes *per se*, or by whether they had to discriminate nested processes with weak distributions of evidence.

Experiment 1: Judging randomness in the frequency of events

As mentioned above, Experiment 1 examined judgments about whether a coin was random (equally likely to produce heads or tails) or systematically biased (towards heads, or towards tails). It investigated whether our nested-process account provided an accurate characterization of the source of errors in people's randomness judgments. In the non-nested condition participants judged whether sequences were biased towards heads or tails for 50 sequences that covered a range of evidence characteristic of biased coins. In the nested condition participants judged whether a coin was fair (random) or biased for 50 sequences that covered a range of evidence characteristic of fair and biased coins. In the matched condition judgments concerned whether a coin was biased towards heads or tails, but the LLR (the nested-process model's measure of the evidence a sequence provided) was used to select 50 sequences that provided exactly as much evidence for a bias to heads/tails as the 50 in the nested condition provided for a fair/biased coin. Although the task differed, the LLRs according to the nested and non-nested models were matched in the nested and matched conditions. If the model captures the statistical difficulty in this randomness judgment, accuracy in the matched condition should be significantly lower than in the non-nested condition, and closer or even equal to that in the nested condition.

Methods

Participants. Participants were 120 undergraduate students (40 in each of three conditions), participating for course credit.

Materials. The 50 sequences in the nested and non-nested condition were chosen to span a range of sequences that would be generated under the nested and non-nested processes. Table 1 shows the distribution of LLRs for the sequences in each condition, as well as example sequences in each range of LLRs, summarized by the number of heads in the sequence. For the nested condition, 50,000 sequences of 40 coin flips were generated by simulating a fair coin (random process) and 50,000 by simulating coins that had $P(\text{heads})$ ranging uniformly from 0% to 100% (systematic process).[5] The 100,000 samples were pooled and ordered by increasing LLR and 50 sequences were selected that covered the range of LLR values by selecting a sequence at every second percentile. A similar process was used for the non-nested condition: 50,000 sequences from a coin with $P(\text{heads}) = 0.3$ and 50,000 from a coin with $P(\text{heads}) = 0.7$ were pooled and ordered by LLR (the evidence for bias towards heads versus

[5] $P(\text{heads})$ for each of the 50,000 sequences was randomly chosen, with all values between 0 and 1 equally likely.

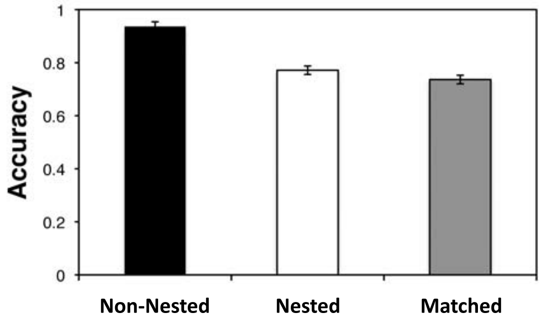


Figure 3. Judgment accuracy as a function of task and evidence in Experiment 1. Error bars represent SEM.

tails) and a sequence selected from every 2nd percentile for a total of 50.

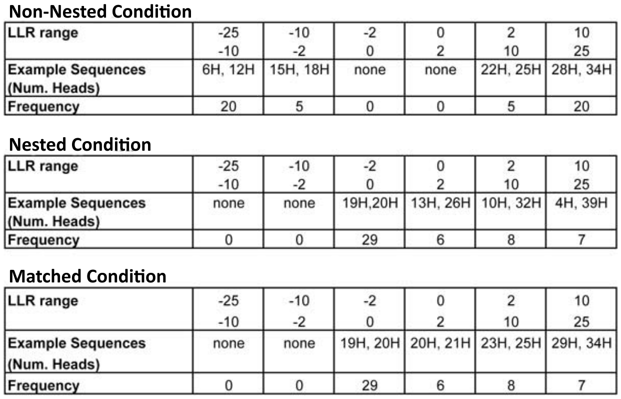
The 50 matched sequences provided the critical test. Participants would judge whether these sequences were biased towards heads or biased towards tails, so the LLR was calculated with respect to the non-nested processes. However, each of the 50 matched sequences was chosen to have a similar LLR to one of the 50 nested sequences. It was not always possible to make the LLRs in the nested and matched condition identical, but sequences were selected to minimize the differences. The sequences in the matched and nested condition were thus matched in the amount of evidence they provided for their respective judgments, but these judgments were about qualitatively different processes.

Procedure. The experiment was administered by computer. Participants in the nested condition were instructed that they would see sequences of coin flips, and that half of these had come from a fair coin that produced heads and tails with probability 50%, and from other half from a coin biased to show heads and tails with some probability other than 50%, with all probabilities being equally likely. For each sequence, they were instructed to decide which process had generated it. Participants in the non-nested and matched condition were instructed that half of the sequences came from (1) a coin that came up heads 30% of the time (tails 70%), and the other half from (2) a coin that came up heads 70% of the time (tails 30%). Participants were given 16 practice trials of just five flips, followed by the actual experiment of 50 trials of 40 flips. Each trial displayed the sequence of heads and tails onscreen, e.g., "HTHTHTHTHHHHHTTHHHHT-THTTTTHHHTTTTHHHTHHT". Responses were made by pressing one of two buttons, with the button-response pairing randomly chosen for each participant.

Results

Accuracy. People's judgment accuracy in each of the three conditions is shown in Figure 3. An accuracy score was constructed for each participant as the proportion of correct inferences out of 50, with an inference scored as correct if the participant chose the process favored by the evidence a sequence provided (its LLR).[6] Accuracy in the non-nested condition was significantly better than in the nested and matched conditions ($t(78) = 6.9, p < 0.001, d = 1.54$; $t(78) = 8.6, p < 0.001, d = 1.87$). However, accuracy in the matched condition did not differ significantly from accuracy in the nested condition ($t(78) = -1.6, p = 0.12, d = -0.30$). When the distribution of evidence for judging randomness and judging direction of bias is equated, people make just as many errors and performance is not significantly different. In fact, accuracy was numerically lower in the matched condition, so any potential differences run counter to the prediction that the model isn't sufficient to capture the difficulty of the task. This provides evidence that the nested-process account accurately characterizes the statistical challenge inherent in this randomness judgment.

Table 1
Distribution of the LLR for sequences used in Experiment 1. Note: The LLR is calculated with respect to the relevant non-nested or nested processes in each condition. The table gives the number of sequences in each condition which have LLRs in a particular range. Example sequences in each LLR range are provided, summarized by the number of heads.



Non-Nested Condition

LLR range	-25 -10	-10 -2	-2 0	0 2	2 10	10 25
Example Sequences (Num. Heads)	6H, 12H	15H, 18H	none	none	22H, 25H	28H, 34H
Frequency	20	5	0	0	5	20

Nested Condition

LLR range	-25 -10	-10 -2	-2 0	0 2	2 10	10 25
Example Sequences (Num. Heads)	none	none	19H,20H	13H, 26H	10H, 32H	4H, 39H
Frequency	0	0	29	6	8	7

Matched Condition

LLR range	-25 -10	-10 -2	-2 0	0 2	2 10	10 25
Example Sequences (Num. Heads)	none	none	19H, 20H	20H, 21H	23H, 25H	29H, 34H
Frequency	0	0	29	6	8	7

Model predictions: Degree of belief. Figure 4 shows the proportion of people choosing h_1 for each of the 50 sequences, as well as the posterior probability of h_1 according to the model's analysis of the judgment – the precise degree of belief in h_1 that is warranted by the evidence the data provide.[7] The sequences are ordered from left to right by increasing LLR. The key pattern illustrated in Figure 4 is that there is a striking quantitative correspondence between the

[6] This is equivalent to the process with higher posterior probability, when the prior probabilities of h_0 and h_1 are equal. Accuracy could also have been evaluated in other ways – such as based on the generating process. We use such an approach in Experiment 2, which has complementary advantages and disadvantages.

[7] The model assumes that h_0 (a random process) and h_1 (a systematic process) are equally likely a priori (the instructions provided to participants also indicate that this is the case) and so the posterior probability depends only on the LLR: it is equal to $\frac{1}{1+e^{-LLR}}$

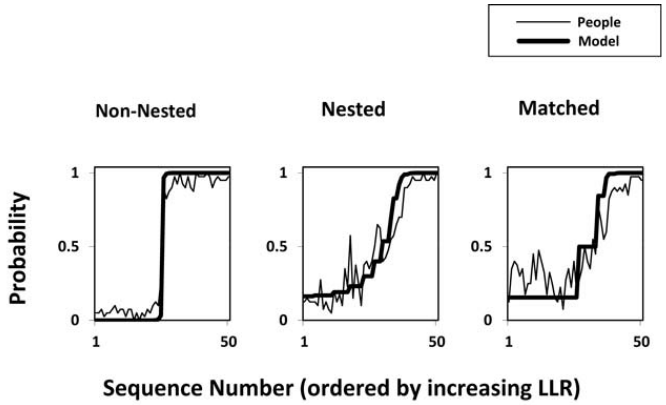


Figure 4. Experiment 1: Human judgments are the proportion of people reporting that a sequence was generated by h_1 and model judgments are the posterior probability of h_1 for a sequence. h_1 represented a systematically biased process for the nested condition and a bias to heads for the symmetric and matched. Sequences are ordered by increasing evidence for h_1.

proportion of people who choose a process and the model's degree of belief in that process, according to the LLR of the data set. The correlations for the non-nested, nested, and matched conditions are, respectively, $r(48) = 0.99$, 0.94, and 0.92, providing compelling evidence that the model accurately captures the statistical difficulty in detecting randomness.

Model predictions: Reaction Time. All reaction time analyses were carried out on data that were first scaled for outliers (reaction times greater than 10 seconds were replaced by a value of 10 s). Reaction time data confirm the pattern of difficulty in judgments: people were faster to make judgments in the non-nested than either the nested ($t(78) = 2.5, p < 0.02, d = 0.56$) or matched ($t(78) = 2.7, p < 0.01, d = 0.60$) conditions, although reaction time for the matched condition was not significantly different from the nested condition ($t(78) = -0.33, p = 0.74, d = -0.07$).

Reaction time was also analyzed as a function of individual sequences (or rather, their LLRs) to obtain detailed model predictions. There was a clear linear relationship between the time people needed to make a judgment about a sequence and the magnitude of the evidence that sequence provided (the size of the LLR). The correlations between the time to make an inference from a sequence and the absolute value of the LLR of the sequence were $r(48) = -0.82$ (non-nested), -0.84 (nested), and -0.75 (matched). The smaller the magnitude of the LLR, the longer the time to make a judgment, the larger the LLR, the quicker an inference was made. The close match between data and model illustrates that the sequences which provide only weak evidence are the sequences that people find inherently difficult to evaluate and spend more time processing.

Discussion

Experiment 1 provided evidence for the nested-process account. Although judgments about a random process (fair coin) were less accurate than similar task judgments abut non-nested processes (head/tail bias), these were due to the weak evidence available rather than people's erroneous intuitions about randomness. The critical matched condition required judgments about non-nested processes, eliminating the role of biases about randomness in erroneous judgments. But it also equated the amount of evidence sequences provided to the evidence available in the nested condition. This eliminated the significant differences, so that the nested and matched condition were equally accurate. Judgments about randomness are more inaccurate than judgments about two kinds of systematic processes not only because they involve reasoning about randomness, but because judgments about randomness are judgments about a nested process.

Across a range of sequences, the proportion of people who judged a random process to be present closely tracked the rational degree of belief an ideal observer would possess based on the statistical evidence available. There was also a close correspondence between the strength of evidence and the difficulty of making a judgment, as measured by reaction time. The results suggest that the assumptions of the model about how processes are mentally represented and related to data provides a good account of participants' difficulty and errors in judging randomness, by closely capturing the uncertainty in the evidence available. In particular, the high correlations between model and data suggest that people are very sensitive to the evidence a sequence provides for a process and are good at judging how likely it is that a particular process generated a sequence.

When a sequence provides strong evidence for a process and the rational degree of belief in that process is high, a high proportion of people choose that process, while when the rational degree of belief is low, few people choose it. Moreover, people's inaccuracy can be understood in terms of closely tracking the rational degree of belief. For example, if the posterior probability of a process is 0.6 and 60% of people choose it, 40% of them are judged to have made an error. However, this error is a consequence of the fundamental uncertainty in the task – even a rational analysis suggests that the data only weakly identify the generating process.

One concern might be whether participants' probability-matching behavior truly provides evidence for the model: If each individual knows that the posterior probability of a process is above 0.5, then perhaps the proportion choosing that process should be 100%, not the posterior probability. However, there are several reasons this may not be true. Even if the evidence available is constant across participants, the particular criterion each uses for a judgment may vary. Also, participants do not necessarily have direct access to a quantitative measure of the evidence in a stimulus, but may process a noisy function of this evidence. As the evidence becomes stronger both an ideal observer's confidence and participants' correct responses increase because they are sensitive to the same statistical information, even if different procedures or processes map this information to a particular judgment. Vulkan (2000) and Shanks, Tunney, and McCarthy (2002) provide further discussion of issues relating to this kind of probability matching.

Experiment 2: Dissociating random processes and nested processes

One reason Experiment 1 provided compelling support for a nested model was that the nested and matched condition were equally accurate *despite* differing in whether participants had to reason about randomness. But a drawback of this difference is that the comparison of the nested to the matched (and non-nested) condition does not isolate being *nested* as a critical feature. Experiment 2 addressed this issue by replicating and extending Experiment 1 in two ways.

The first was that the nested judgment was compared to a judgment that was both non-nested *and* required reasoning about a random process, providing a closer match. This *random non-nested* condition required discriminating a random coin ($P(\text{heads}) = 0.5$) from a biased coin that produced heads 80% of the time. Although this condition also requires detecting a random process, the nested model predicts a more informative distribution of evidence and higher accuracy since the judgment about randomness is not *nested*. The matched condition was similarly adapted to provide a more direct comparison to the nested condition by using the random non-nested judgment task, but statistically matching the evidence to that in the nested condition, which we now label the *random nested* condition.

For the second extension we included two conditions that allowed us to *independently* manipulate whether participants made judgments about random (vs. only systematic) processes, and whether the judgments were nested (vs. non-nested). The *systematic non-nested* condition did not require reasoning about a random process, but was chosen to be statistically similar to the random non-nested condition. It required discrimination of a process with $P(\text{heads}) = 0.4$ from one with $P(\text{heads}) = 0.7$. The *systematic nested* condition was statistically similar to the random nested condition. It required evaluating whether a sequence was generated by a systematically biased coin with $P(\text{heads}) = 0.4$ or a biased coin with $P(\text{heads})$ between 0 and 1.

This constitutes a two (judgment: requires vs. does not require consideration of a random process) by two (statistical structure: nested vs. non-nested) design. Anested-process account predicts a main effect of statistical structure – where accuracy is lower for nested than non-nested processes – but no effect of whether the judgment involves consideration of a random process. Alternatively, if the involvement of random processes is what makes a task hard, accuracy should be lower whenever people have to use their concept of randomness or apply a heuristic in evaluating a random process. Finally, if the statistical structure of the task is irrelevant, we should see no difference between the nested and non-nested judgments.

Two further changes were made to complement Experiment 1. To more directly target participants' intuitions about randomness, they were instructed to judge whether a sequence reflected a random coin. Experiment 1 framed the task as identifying whether the had a 50% probability of heads, which was logically equivalent but may not have directly tapped intuitions about randomness. Also, Experiment 1 selected sequences with a broadly representative range of LLRs and so had to use the ideal observer model to assess accuracy. Experiment 2 chose the sequences to represent the distribution that was most likely under each process and used this information in scoring accurate responses. Using both of these methods for selecting sequences and scoring accuracy ensures that our findings are not an artifact of any particular method.

Methods

Participants. Participants were 110 members of the Amazon Mechanical Turk community who participated online for monetary compensation and 90 undergraduate students who participated for course credit (40 in each of five conditions, with participants from the different sources spread evenly across the conditions).

Materials. The systematic non-nested ($P(\text{heads}) = 0.4$ vs. 0.7), random non-nested ($P(\text{heads}) = 0.5$ vs. 0.8), systematic nested ($P(\text{heads}) = 0.4$ vs. $[0-1]$), and random nested ($P(\text{heads}) = 0.5$ vs. $[0-1]$) conditions each presented 50 sequences of 40 coin flips, 25 from each process. 25 sequences that represented those expected from each process were selected by multiplying 25 by the probability distribution over sequences for each process. The answer was used to choose how many sequences had a particular number of heads. For example, if $P(\text{heads}) = 0.5$ the probability of a sequence with 20 heads is 0.125 and so there were three sequences with 20 heads ($0.125 \times 25 = 3.125$, the product was rounded). For the matched condition, the 50 sequences were selected so that the LLR with respect to the random non-nested judgment ($P(\text{heads})$ of 0.5 vs. 0.8) was as similar as possible to the LLRs in the random nested condition.

Procedure. Participants were informed that they would see sequences of heads and tails that were generated by different processes and that they would judge what the generating process was. For each condition, they were informed what the relevant processes were and told that half of the coins came from each process. For example, in the random nested condition they were told that half of the sequences came from a coin that is random – has 50% probability of heads and half from a coin that has an 80% probability of heads. Each trial displayed the sequence onscreen, e.g., "HHTHTHTHTHHHHTTHHHHTTHTTTTHHHTTTTHHHTHHT". The order of the flips in a sequence was randomized on each presentation. Responses were made on the keyboard. To familiarize participants with the task they had a practice phase of making judgments about 16 sequences of just five flips. The actual experiment required judgments for 50 sequences of 20 flips.

Results & Discussion

Accuracy was calculated in two ways. First, as in Experiment 1 an inference was scored as correct if the participant chose the process favored by the evidence a sequence provided (its LLR). Second, an inference was scored as correct if it corresponded to the process whose distribution was used to select the coin. Both of these measures gave the same pattern of results, and to be consistent throughout the paper we report the first. Figure 5 shows accuracy for all five conditions: systematic non-nested, random non-nested, systematic nested, random nested, matched.

Comparison of random non-nested, random nested, and matched judgments. Although both conditions involved a judgment about a random process, accuracy was significantly lower in the random nested than the random non-nested condition ($t(78) = -4.67, p < 0.001, d = 1.04$). This reflects the particular challenge of discriminating nested processes. To test whether this was due to weaker evidence in the random nested condition, the matched condition judged whether a sequence was from a coin with P(heads) = 50% or 80%, but only for sequences with LLRs matched to those in the random nested condition. Accuracy in the matched condition was also significantly lower than the random non-nested ($t(78) = -4.61, p < 0.001, d = 1.03$), but did not differ significantly from the random nested condition ($t(78) = 0.09, p = 0.93, d = 0.02$). This replicates the finding from Experiment 1 that a weaker distribution of evidence was responsible for errors in randomness judgment, which is underscored by the better accuracy in reasoning about a random process when it was not nested.

Judgment as a function of whether a process is random and/or nested. Accuracy can also be analyzed as a function of two key independent variables: whether a judgment condition involved a random vs. systematic process, and whether it involved a nested vs. non-nested process. This generates the four conditions: systematic non-nested, random non-nested, systematic nested, and random nested. Accuracy (see Figure 5) was analyzed in a two (random vs. systematic) by two (nested vs. non-nested) ANOVA. Judgments that involved nested processes were significantly less accurate than judgments about non-nested processes ($F(1,195) = 65.22, p < 0.001$). However, there was no effect of whether a judgment involved reasoning about a random or systematic process ($F(1,195) < 1, p = 0.94$). In support of the effect of nested processes on errors in randomness judgment, this factor had a more substantial effect on reasoning errors than whether or not people had to consider a random process.

The interaction from this ANOVA was not significant ($F(1,195) < 3.56, p = 0.06$). Accuracy in the systematic non-nested condition did not differ significantly from accuracy in the random non-nested condition ($t(78) = -0.28, p = 0.78, d = -0.06$). Accuracy in the systematic nested condition did not differ significantly from the random non-nested condition ($t(78) = 1.79, p = 0.08, d = 0.4$), and if anything the trend was for it to be lower. The proportion of people reporting that a particular process generated a sequence was closely predicted by the posterior probability of that process under the model. The correlations for each condition were systematic non-nested, $r(48) = 0.96$, systematic nested, $r(48) = 0.84$, random non-nested, $r(48) = 0.95$, random nested, $r(48) = 0.84$, and matched $r(48) = 0.86$. Reaction time data was not analyzed because many participants conducted the experiment online, precenting accurate measurement of reactiontimes for all participants. Across a range of tasks, the ideal observer analysis provided a compelling account of people's judgments, in terms of the evidence available in evaluating nested, random and systematic processes.

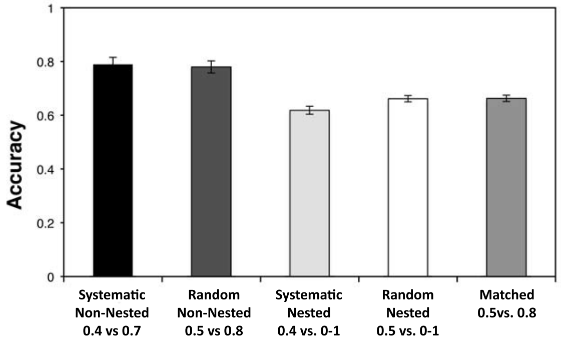


Figure 5. Judgment accuracy as a function of task and evidence in Experiment 2. Error bars represent SEM.

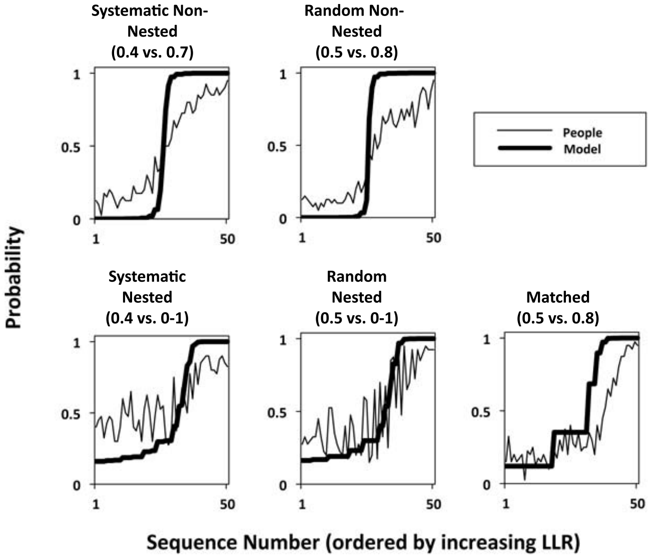


Figure 6. Experiment 2: Proportion of people reporting that a sequence was generated by h_1 and posterior probability of h_1 under the model. h_1 represented a systematically biased process for the nested conditions and a bias to heads for the symmetric and matched conditions. Sequences are ordered by increasing evidence for h_1.

Experiment 3A: Evaluating randomness vs. sequential dependence

Experiments 1 and 2 showed how a nested process restricts the amount of evidence available and thus causes judgment errors. Experiment 3 extended the nested-process account to judgments about randomness and systematicity in the context of sequential dependence. The judgment was whether successive coin flips were random in being independent of each other or exhibited systematic sequential dependence: the probability of a repetition (and alternation) was not 50%. The conceptions of randomness and reasoning strategies people use for evaluating sequential independence may differ from previous experiments, but still share the key statistical feature of evaluating a nested process. Another critical feature of this judgment is that there is ample empirical evidence that errors are caused by misleading intuitions about dependence. People demonstrate an *alternation bias* whereby they believe that sequences with many alternations (e.g., alternating from heads to tails) are more random than sequences with many repetitions (e.g., repeating heads or tails), and that repetitions are more likely to reflect systematic processes (Falk & Konold, 1997; Bar-Hillel & Wagenaar, 1993). Judgment errors may therefore be most parsimoniously explained by biases in reasoning. Extending the nested-process model to this context can shed light on the model's utility in integrating both inherent statistical limitations and flawed reasoning.

In fact, as shown in our earlier presentation of the model, a random sequentially independent process is nested in the set of systematic processes that have a bias to repeat or alternate. The nested-process account predicts that judgment should be impaired by *both* misconceptions *and* the weak evidence available. On the other hand, it may be that the presence of biased intuitions swamps any contribution of statistical limitations, especially if judgment is already poor. Using inaccurate representations to reason about dependence could also people's sensitivity to the evidence provided, such as if resolving inconsistencies reduces processing resources.

To test whether statistical limitations make a contribution to errors above and beyond biases, we aimed to capture people's alternation bias in an augmented *biased* model of the judgment. This analysis embraces the fact that an ideal observer may entertain misleading hypotheses that do not match the structure of the world, but still be sensitive to the evidence observations provide for those hypotheses.

The biased model. People's alternation bias is illustrated in Figure 7 (data from Falk and Konold, 1997, Experiment 3). Apparent randomness ratings are plotted as a function of how likely the sequence is to alternate. The alternation bias is obvious when human judgments are compared to our first model, which from this point on we label the *uniform* model because it assumes that all systematic processes are equally likely. The model ratings of randomness shown in Figure 7 were computed by evaluating the LLR a sequence provides and scaling it to the same range as human judgments. Although the uniform model captures the general trend, it fails to capture human ratings of alternating sequences as more random than repeating sequences. The model does not accurately capture the processes people represent and reason about.

The key to the novel *biased* model was replacing the assumption that all systematic processes were equally likely (a uniform distribution over P(repetition)) with an assumption that systematic processes were more likely to be repeating than to be alternating. As we consider in the General Discussion, different approaches could be taken, but our goal was simply to capture the bias accurately enough to test the key prediction about nested processes. The assumption that systematic processes are *more* likely to be repeating than alternating captured by defining a *beta* distribution rather than uniform distribution over P(repetition). The mathematical details of how the parameters of this distribution were selected are presented in Appendices A and C, but in Experiment 3A they were chosen to capture the magnitude of the alternation bias in data from Falk and Konold's (1997) Experiment 3. In Experiment 3B the parameters were then chosen to capture the alternation bias in Experiment 3A, so that our findings would not be an artifact of a specific parameter choice. Figure 7 shows that the biased model better captures people's judgments about the relative randomness of repetitions and alternations.

Even when the alternation bias is incorporated into the model, the random process is still nested in a range of systematic processes. Replicating our ideal observer analysis with the biased instead of uniform model produces the same results: randomly generated data have weaker LLRs and should lead to more errors, in addition to those caused by the alternation bias. As in previous experiments, a nested, matched and non-nested condition was compared. However, in contrast to previous experiments, participants in the nested condition were simply informed that the coin came from a "random" or "non-random" process and given *no* information about these processes. Because the judgment relies only on people's intuitions about what "random" and "non-random" means, this provided a strong test of whether our nested-process account truly characterizes the challenges in human reasoning about randomness.

Methods

Participants. Participants were 120 undergraduate students (40 in each of three conditions) who received course credit.

Materials. Sequences were selected using a similar method to Experiment 1. However, the number of flips was reduced to 20 and all sequences used exactly 10 heads and 10 tails, consistent with previous research (Falk & Konold, 1997).

For nested sequences, 50,000 sequences were generated by simulating a random coin with independent flips (P(repetition) = 0.5). Another 50,000 sequences were generated by simulating a coin that was biased to repetition or

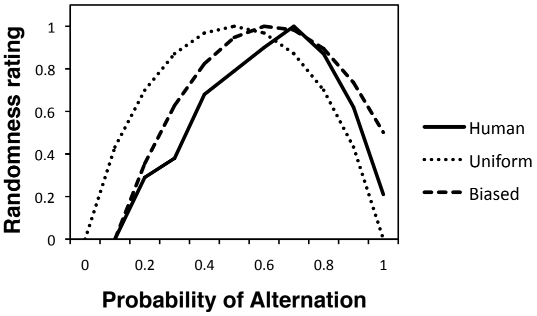


Figure 7. Human and model randomness ratings for Falk and Konold (1997). Human: Human rating. Uniform: Rating for model in which repetitions and alternations are judged equally systematic. Biased: Rating from the model biased to judge repetitions as more systematic than alternations.

alternation (P(repetition) ranged uniformly from 0 to 1).[8] The LLR of each sequence was computed under the biased model, all sequences were pooled and ordered by increasing LLR, and 50 sequences selected by choosing one at each 2nd percentile.

For non-nested sequences, 50,000 sequences were generated by simulating a coin biased to repeat (P(repetition) ranged uniformly from 0.5 to 1) and 50,000 by simulating a coin biased to alternate (P(repetition) ranged uniformly from 0 to 0.5). The LLR of each sequence was computed (relative to the non-nested processes of a bias to repetition or alternation), the sequences pooled and ordered, and 50 which spanned the range of LLRs were selected.

For matched sequences, the LLRs in the nested condition were used to select *two* sets of 25 matched sequences. In matching the LLRs of the first set of 25 matched sequences, positive LLRs provided evidence for repetition, while in the second set positive LLRs provided evidence for alternation. The distribution of the LLRs for the nested sequences was not symmetric around zero (ranging from -0.8 to +4.0), so this control ensured that the matched sequences provided the same overall amount of evidence for repetition and alternation, guarding against possible asymmetries in judgment. The 25 LLRs used in the matched condition still spanned the full range of evidence: they were obtained by averaging every two successive LLRs in the nested condition (ie. the LLRs of the 1st and 2nd sequences, 3rd and 4th, and so on up to the 49th and 50th).

Procedure. Participants were informed that they would see sequences of heads and tails that were generated by different computer simulated processes, and that their job would be to infer what process was responsible for generating each sequence. In the non-nested and matched condition participants were instructed that about half the sequences were generated by computer simulations of a coin that tends to *repeat* its flips (go from heads to heads or tails to tails) and the other half by simulations of a coin that tends to *change* its flips (go from heads to tails or tails to heads). In the nested condition participants were simply told that half the sequences were generated by computer simulations of a *random* process and that half were generated by simulations of a *non-random* process. Participants received a practice phase where they made judgments about 16 sequences of just 5 flips, to familiarize them with the task. They then provided judgments for 50 sequences of 20 flips. Each trial displayed the sequence of heads and tails onscreen (e.g., HHTHTHTHTHHHHTTHH-HHTT).

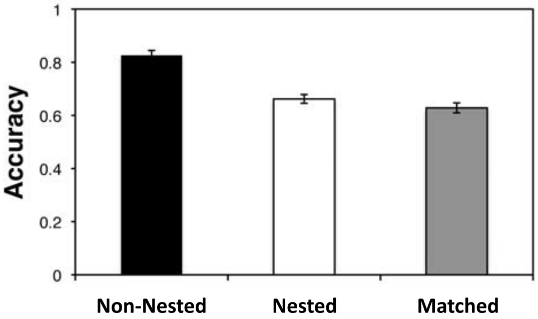


Figure 8. Judgment accuracy as a function of task and evidence in Experiment 3A. Error bars represent SEM.

Results & Discussion

Accuracy for each condition is shown in Figure 8. As in Experiment 1, accuracy was significantly higher in the non-nested condition than the nested and matched condition ($t(78) = 6.61, p < 0.0001, d = 1.48$; $t(78) = 7.48, p < 0.0001, d = 1.67$). However, there was no significant difference between the matched and nested conditions ($t(78) = -1.35, p = 0.18, d = -0.3$). Once the evidence that the sequence provides was equated to that of sequences in the nested condition, the difficulty of judging whether a sequence was biased to alternate and repeat was not significantly different from judging whether it was random or not. People face a double challenge in randomness judgments. Not only do misconceptions like the alternation bias reduce accuracy, but the inherent statistical limitations on the evidence available for a nested random process also generate errors.

Figure 9 shows the posterior probability of h_1 under the model and the proportion of participants choosing h_1, across all three conditions. The proportion of participants choosing the hypothesis closely tracked the degree of an optimal reasoner's belief in that hypothesis. The correlations between

[8] Although the model assumes the representation of a systematic process is biased towards repetitions, the generation of systematic sequences did not reflect this bias to ensure that the sequences would be representative of *actual* random and systematic processes. However, the biased model was used to compute the LLR and determine how much evidence a sequence provided for a random process.

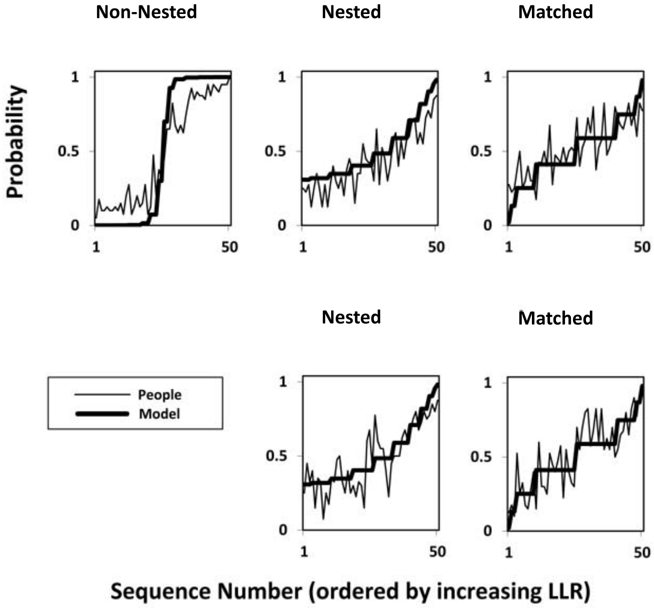


Figure 9. Experiment 3A and 3B: Proportion of people reporting that a sequence was generated by h_1 and posterior probability of h_1 under the model. h_1 represented a sequentially dependent process biased to repeat or alternate for the nested condition, and a bias to repeat flips for the symmetric and matched conditions. Sequences are ordered by increasing evidence for h_1.

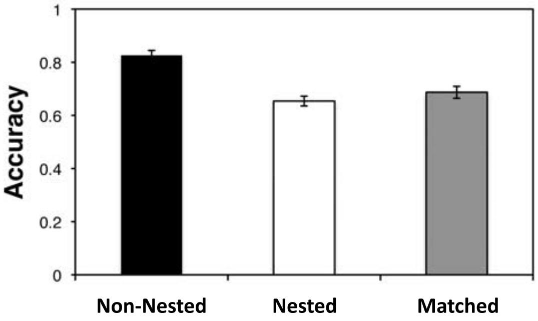


Figure 10. Judgment accuracy as a function of task and evidence in Experiment 3B. Error bars represent SEM.

the model predictions and human judgments were $r(48) = 0.96$ (non-nested), 0.87 (nested), and 0.79 (matched). People's uncertainty and errors closely tracked the rational degree of belief a reasoner should have based on the evidence a sequence provided. This is particularly noteworthy because people were not told the nature of the random and systematic processes and had to rely on their intuitions about "random" and "non-random" process. There were no significant differences across conditions in the time to make judgments about sequences (all ps > 0.64, all ds< 0.10). The correlations between the absolute value of the LLR and reaction times were $r(48) = -0.52$ (non-nested), -0.68 (nested), and -0.23 (matched). Reaction time may have been less informative because the range of LLRs was smaller than previous experiments.

Experiment 3B: Gaining a closer match to human biases

Experiment 3A assumed that the alternation bias was similar to that in Falk and Konold's (1997) experiment, but these populations and tasks may differ in significant ways. An informative comparison relies on the model representing similar hypotheses to people in a particular task. Our goal in Experiment 3B was to ensure that the results were not dependent on the particular model and parameters used to capture the alternation bias in Experiment 3A. Just as Experiment 3A constructed a biased model to account for the alternation bias in Falk & Konold (1997), Experiment 3B replicated Experiment 3A using a model constructed to account for the alternation bias in Experiment 3A, by inferring the parameters that capture the alternation bias from participants' randomness judgments. The same procedure was used as in Experiment 3A's construction of the biased model, and the details are reported in Appendix C. While the new parameters differed from those in Experiment 3A, the size of the alternation bias was similar to that found by Falk and Konold (1997). Changing the model parameters did not influence the non-nested condition, so only the nested and matched conditions were replicated.

Methods

Participants. Participants were 120 undergraduate students (40 in each condition) who received course credit.

Materials. The procedure used to generate sequences was identical to that of Experiment 2, except that new parameters were used for the *biased* model.

Procedure. The procedure was identical to Experiment 2.

Results & Discussion

The results replicated the findings in Experiment 2. Figure 10 shows accuracy across conditions. Accuracy in the non-nested condition was significantly better than in either the nested ($t(78) = 6.56, p < 0.0001, d = 1.47$) or matched condition ($t(78) = 4.74, p < 0.0001, d = 1.06$), although there was no difference in accuracy between the matched and nested conditions ($t(78) = 1.13, p = 0.26, d = 0.25$). Figure 9 shows the posterior probability of h_1 under the model and the proportion of people choosing h_1. The correlations between human judgments and model predictions were $r(48) = 0.83$ in the nested and 0.85 in the matched condition. Reaction times did not differ between the non-nested and nested conditions ($t(78) = 0.03, p = 0.98, d = 0.00$), but judgments took significantly longer in the matched condition than either the non-nested ($t(78) = -3.45, p < 0.001, d = -0.77$) or nested ($t(78) = -3.51, p < 0.001, d = -0.79$) condition. The correlation between reaction time and LLR was $r(48) =$

-0.52 (non-nested), -0.73 (nested), and -0.16 (matched). Overall, Experiment 3B replicated the key findings of Experiment 3A, showing that its findings were not restricted to the particular parameters used, and that the analysis provides a reasonable characterization of how people represented the processes.

General Discussion

We presented a *nested-process* model to characterize how statistical features of randomness judgment generate errors and report three experiments providing evidence for this model. Specifically, we examined the task of evaluating whether a sequence of binary outcomes was generated by a random process (in that either outcome was equally likely) or a systematic process (there was a bias for one outcome to be more likely than the other). Of the many real-world tasks that share this formal structure, we examined whether sequences of coin flips were random versus biased towards heads/tails, and random versus biased towards repetitions/alterations. The key contribution was a statistical model of these random and systematic processes (equiprobable versus biased outcomes). The strength of evidence that sequences provided for random versus systematic processes was quantified using the statistical measure of the Log-Likelihood Ratio (LLR), and showed that randomly generated data sets overall provided weaker and less diagnostic evidence for a random process, than systematically generated data sets provided for a systematic process. Our nested-process account further explains why this weaker distribution of evidence is found – because the random process is *nested* inside the set of systematic processes.

To illustrate with an example, consider evaluating whether a sequence of coin flips is generated by a random process (heads and tails equally likely) versus a systematic process (biased towards heads/tails). The random process can be modeled by $P(\text{heads})$ of 50% , which is nested within the set of systematic processes, for which $P(\text{heads})$ ranges between 0% and 100%. This nested relationship means that randomly generated sequences (from a coin with $P(\text{heads})$ of 50%) could also have been generated by systematically biased coins (e.g.with $P(\text{heads})$ of 40%, or 55%), which results in randomly generated sequences providing only weak evidence for randomness and having small LLRs. Achieving a high rate of correct detections of systematic processes can necessitate a high rate of false alarms – the quintessential error of attributing randomly generated data to a systematic process.

Experiments 1 and 2 examined people's accuracy in evaluating whether sequences were generated by a coin that was random (heads/tails equally likely) versus a coin systematically biased towards heads/tails. Accuracy of judgments was compared across *nested*, *non-nested*, and *matched* conditions. The nested condition required discriminating a random (and therefore nested) process from a systematically biased process. The non-nested condition required discriminating two systematic (and therefore non-nested) processes. While more errors were made in the nested than the non-nested conditions, our model suggested that this was due to the stronger evidence provided by sequences in the non-nested condition. We obtained empirical evidence for this conclusion by constructing the critical matched condition, in which the judgment task involved systematic processes (matching the non-nested condition) but the sequences presented provided similar evidence and had similar LLRs to the nested condition. Judgment errors in the matched condition did not differ significantly from the nested condition but were far greater than the non-nested condition. Once the evidence available in judgments about systematic processes was equated, accuracy was just as bad. Experiment 2 also independently manipulated whether judgments required reasoning about a random process or not, and reasoning about a nested process or not. The results suggested that whether or not a judgment concerns nested processes was more important in predicting errors, than whether or not it required reasoning about a random process.

Experiment 3 generalized the key findings to judgments about whether a coin was random (in that successive flips were independent) vs. systematically biased (towards repetition or alternation). Both Experiment 2 and Experiment 3 used language that identified a coin with a 50% probability of heads as a *random* coin, to rule out concerns that the results depended on it being easier to reason about the former than the latter. Overall, across a range of data sets, people's reaction times to make a decision and the particular random or systematic process they identified was closely predicted by the ideal observer model's quantification of the evidence available.

In the remainder of the paper, we revisit some of the assumptions behind our analysis and discuss its limitations, discuss the relationship between our rational model and biases in judgment, and consider future research and practical implications suggested by our findings.

Assumptions, limitations and extensions of the nested-process account

A quick reading might suggest that our nested-process account makes an obvious point about randomness judgment. While it may be intuitive that identifying a nested process is difficult, the proposal that this is a key feature of random processes is a novel one. Despite awareness of the mathematical difficulty in evaluating randomness (e.g., Lopes, 1982), this paper is novel in proposing a specific formal model, deriving and quantifying its implications for judgment using a measure of evidence like the LLR, and empirically testing whether the model explains people's errors.

Another worry could be that the comparison to non-nested processes was reliant on artificially constructing easier judgments (e.g., discriminating coins with $P(\text{heads})$ of 30% and 70%). However, it should be noted that our ideal observer analysis shows that the key result – an asymmetric, weak distribution of evidence – does not depend on the specific parameters so much as whether the processes are nested or not. We explore a range of parameter choices by examining judgment about non-nested systematic processes that are represented by a single parameter (Experiment 1) or an entire interval (Experiment 3), symmetric (Experiment 1) or skewed towards alternation (Experiment 3). Moreover, Experiment 2 replicated Experiments 1 and 3 even when comparing nested and non-nested judgments that *both* involved random processes, and found that errors were primarily a consequence of whether a process was nested rather than random, even when the statistical features of judgments about random and systematic processes were closely matched (e.g. judging whether P(heads) = 50% or P(heads) = 40% vs. P(heads) ranging between 0% and 100%).

The modeling assumption that people represent systematic bias in the occurrence of events as uniformly distributed (evaluating fair vs. biased coins, Experiments 1 and 2) and systematic dependencies as more likely to repeat than alternative (Experiment 3) matched the current experimental results, but these and other assumptions of the current modeling framework can certainly be improved. For example, future modeling and empirical work could more precisely characterize the nature of people's beliefs about the distribution of systematic processes across different contexts.

The roles of rationality and biases in randomness judgments

In using ideal observer models to understand how people evaluate randomness, we are not making a strong claim that people are rational, or arguing that biases are not involved in randomness perception. Rather, we use these rational models as the basis for the claim that there may be factors that combine with biases to make the identification of random processes especially challenging. The *biased* model used in Experiments 3A and 3B demonstrated that an ideal observer analysis may provide an invaluable tool in investigating which aspects of judgment reflect biases and which stem from inherent statistical challenges. In this model, we used existing results concerning the kinds of biases that people have about randomness to inform our assumptions about what form systematic processes might take, so that we could evaluate what an ideal observer with these beliefs would infer and what challenges they would face.

Future modeling work may also identify ways in which particular heuristics and biases have developed to mitigate statistical challenges such as inherently weak evidence. If a learner benefits greatly from discovering true systematicity and pays little cost for misclassifying a random process as systematic, the rational consequence of a cost-benefit utility analysis may be the heuristic use of a liberal criterion that correctly classifies most systematic processes but also misclassifies many random processes as systematic. The human bias to "irrationally see meaning in randomness" may be an adaptive strategy in general that overcomes statistical limitations on evidence, but in isolated judgments has the surface appearance of an irrational phenomenon.

Our formal analyses focused on the fact that the data generated from nested processes provides inherently weak evidence, whether or not the prior probabilities of particular processes varied. This leaves open a number of interesting questions about the effect of manipulating prior probabilities or beliefs. One possibility is that the accuracy of randomness judgments may be especially jeopardized when people have strong and misleading beliefs in systematicity, for example, believing a particular causal relationship exists between an unproven medicinal supplement and health. Even for a completely rational agent, changing strong prior beliefs requires that the data provide strong evidence – which is precisely what randomly generated data do not provide.

Weak evidence, processing limitations, and improving judgment

The statistical properties of the problem of detecting random processes may have implications for other aspects of people's reasoning about chance. Some research on illusory correlation (Redelmeier & Tversky, 1996; Jennings, Amabile, & Ross, 1982) proposes that people erroneously detect structure by selectively attending to the subset of available data which provide evidence for structure and ignoring the data that provide evidence for randomness. It may be that processing limitations mean that people are not able to utilize the many observations relevant to computing or inferring a correlation.

If this forces people to consider only a subset of the data, a rational solution would be to utilize the diagnostic data that provide the most evidence (data points with large LLRs). In many contexts (such as inferences about non-nested processes) the same inference will be reached more quickly and with less computation than using the entire data set. But a randomly generated data set will contain a large amount of weak evidence for randomness and (by chance) a small amount of stronger evidence for structure. While considering *all* observations might provide evidence for randomness, selectively attending to the elements of a randomly generated data set that provide strong evidence (which are data points providing evidence for a systematic process, given that most data points provide only weak evidence for a random process) would lead to inferring a systematic process. This problem would be further compounded if people had any prior reason to believe a systematic process was present. Future work can manipulate the distribution of evidence across samples to investigate this possibility.

The proposal that evaluating randomness is hard because the evidence available is inherently limited suggests a different approach to improving judgment, taking a different tack from attempts to revise misconceptions or biases. One basic means of improving inferences about the presence or absence of randomness is presenting large amounts of data or organizing it such that it can be readily processed. This should make it easier for people to accumulate many weak pieces of evidence for a random process. Calibrating prior beliefs towards expecting random processes and increasing skepticism about the presence of systematic processes may also be a useful prescription. Restricting the breadth of the systematic processes under consideration could also aid judgment: stronger evidence for random processes can be obtained if the alternative hypotheses specify only very strongly systematic processes (e.g., deterministic causal relationships, see Schulz & Sommerville, 2006; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008) or processes that display just a particular form of systematicity (e.g. a bias towards one value of a binary outcome, but not the other). These are all novel and promising directions for future research.

Conclusion

People make numerous errors in evaluating whether observations reflect random processes or underlying systematicity. Many of these errors are due to misconceptions and biases in reasoning about randomness, but a further challenge is the mathematical difficulty of detecting a random process. We presented a nested-process account to characterize the inherent statistical challenge in detecting randomness. Ideal observer analyses simulated using computational models show how a random process is a special case of a systematic process one with no systematicity, so that it is nested or contained in a range of systematic processes. The models demonstrated how this means that even truly randomly generated is still likely to come from a systematic process. This imposes statistical limitations on the evidence the data can provide for a random process, and impairs judgment. Three experiments provided compelling evidence for our account's predictions about human judgments, showing that the weak evidence available in evaluating a nested process plays a substantial role in producing errors. In fact, in our experiments, the strength of this evidence had a greater effect on judgment accuracy than whether or not people had to reason about a random process. By showing how some challenges humans face in detecting randomness are shared with ideal statistical reasoners, we provide a more comprehensive account of why people can be so bad at detecting randomness.

References

Alter, A., & Oppenheimer, D. (2006). From a fixation on sports to an exploration of mechanism: The past, present, and future of hot hand research. *Thinking & Reasoning*, *12*(4), 431–444.

Anderson, J. (1990). *The adaptive character of thought.* Lawrence Erlbaum.

Bar-Hillel, M., & Wagenaar, W. (1993). The perception of randomness. *A handbook for data analysis in the behavioral sciences: Methodological issues*, 369–393.

Boas, M. L. (1983). *Mathematical methods in the physical sciences* (2nd ed.). New York: Wiley.

Chapman, L., & Chapman, J. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*(3), 193–204.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301–318.

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Griffiths, T. L., & Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory.

Hamilton, D. (1981). Illusory correlation as a basis for stereotyping. In D. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 115–144). Hillsdale, NJ: Lawrence Erlbaum.

Jennings, D., Amabile, T., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*, *3*(3), 430–54.

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

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*(3), 263–269.

Lopes, L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(6), 626.

Lopes, L., & Oden, G. (1987). Distinguishing between random and nonrandom events. *J. Exp. Psych.: Learning, Memory, and Cognition*, *13*, 392–400.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P., & Holyoak, K. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984.

Nickerson, R. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330–356.

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

Rapoport, A., & Budescu, D. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, *121*(3), 352–363.

Redelmeier, D., & Tversky, A. (1996). On the Belief that Arthritis Pain is Related to the Weather. *Proceedings of the National Academy of Sciences*, *93*(7), 2895–2896.

Schulz, L., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child development*, *77*(2), 427–442.

Shanks, D., Tunney, R., & McCarthy, J. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of th e Cognitive Science Society* (pp. 1036–1041).

Tune, G. S. (1964). A brief survey of variables that influence random generation. *Perceptual and Motor Skills*, *18*(3), 705–710.

Vulkan, N. (2000). An economists perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118.

Wagenaar, W. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*(1), 65–72.

Appendix A
Models used in comparing nested and non-nested processes

Uniform model of randomness judgment

The number of outcomes (e.g., heads or repetitions) in a given sequence follows a binomial distribution that depends on n, the number of potential outcomes, and the probability of an outcome, p. The nested hypotheses for a typical randomness judgment are h_0: $p_1 = 0.5$ and h_1: $p_2 \sim \text{Uniform}(0,1)$. If k is the number of times the outcome occurs in a sequence, the log likelihood ratio is

$$\log \frac{P(d|h_1)}{P(d|h_0)} = \log \frac{\text{Beta}(k+1, n-k+1)}{(p_1)^n}$$

where the beta function $\text{Beta}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ (Boas, 1983).

$P(d|h_0)$ is simply the likelihood of the sequence d with k heads under a binomial distribution, being $(p_1)^k(1-p_1)^{n-k}$. $P(d|h_1)$ uses the likelihood under a binomial, but must integrate this likelihood over the uniform distribution on p_2, and so is derived as follows:

$$\begin{aligned} P(d|h_1) &= \int_0^1 (p_2)^k(1-p_2)^{n-k}dp_2 \\ &= \text{Beta}(k+1, n-k+1) \end{aligned}$$

by the definition of the beta function.

These derivations were also used to model the judgment in Experiment 1 about whether a coin had $P(\text{heads}) = 0.4$ vs. some other bias.

Non-nested judgments about point processes

Representing $P(\text{outcome})$ as p, judgments about the direction of systematic bias (heads vs. tails, repetition vs. alternation) can represented as the non-nested hypotheses h_0: $p_1 = 0.3$ and h_1: $p_2 = 0.7$. The LLR is

$$\log \frac{P(d|h_1)}{P(d|h_0)} = \log \frac{(p_2)^k(1-p_2)^{n-k}}{(p_1)^k(1-p_1)^{n-k}}$$

This derivation was also used to model judgments in Experiment 2 about whether $P(\text{heads})$ was 0.5 vs. 0.8, or 0.4 vs. 0.7.

Non-nested judgments about processes over an interval

Judgments about the direction of systematic bias can also be represented as the non-nested hypotheses $h_0 : p \sim \text{Uniform(0,0.5)}$ and $h_1 : p \sim \text{Uniform}(0.5, 1)$. The LLR can be derived similarly to that for the uniform distribution from 0 to 1, and is

$$\log \frac{P(d|h_1)}{P(d|h_0)} = \log \frac{\text{Beta}(k+1, n-k+1) - \text{Beta}_{0.5}(k+1, n-k+1)}{\text{Beta}_{0.5}(k+1, n-k+1)}$$

where $\text{Beta}_{0.5}$ is defined as

$$\text{Beta}_{0.5}(x,y) = \int_0^{0.5} t^{x-1}(1-t)^{y-1}dt$$

and is the incomplete beta function evaluated at 0.5.

Biased model of randomness judgmentl

Let p represent $P(\text{repetition})$. The alternation bias was captured through changing model assumptions about the distribution of systematic processes. The uniform distribution over p in the uniform model was replaced by the more general beta distribution. The beta distribution is defined by two parameters, α and β, with $P(p) \propto (1-p)^{\alpha-1}p^{\beta-1}$. These parameters have a natural interpretation as representing expectations based on prior experience: α can be interpreted as the number of prior observations of alternations and β as the number of prior observations of repetitions. For example, when α and β are both 1, $p \sim \text{Beta}(1, 1)$ is identical to the uniform distribution assumed in the uniform model, reflecting maximal uncertainty about which processes are likely.

When β is greater than α, the model is biased to expect that systematic processes are more likely to be repetitions than alternations (p greater than 0.5 are more likely), while the reverse is true when α is greater than β. The alternation bias can therefore be modeled by a beta distribution with β larger than α: repetitions will be more diagnostic of systematic processes and alternations thus more diagnostic of a random process. Appendix C explains how these parameters were fit to the alternation bias in human data in order to carry out Experiments 3A and 3B.

The nested hypotheses were represented as h_0: $p = 0.5$ and h_1: $p \sim \text{Beta}(\alpha, \beta)$. The LLR is:

$$\log \frac{P(d|h_1)}{P(d|h_0)} = \log \frac{\frac{\text{Beta}(\alpha+k, \beta+n-k)}{\text{Beta}(\alpha,\beta)}}{(0.5)^n}$$

Using the beta probability distribution over p, the numerator was derived by

$$\begin{aligned} P(d|h_1) &= \int_0^1 ((p)^k(1-p)^{n-k})(\frac{(p)^{\alpha-1}(1-p)^{\beta-1}}{\text{Beta}(\alpha,\beta)})dp \\ &= \frac{\int_0^1 (p)^{\alpha+k-1}(1-p)^{\beta+n-k-1}dp}{\text{Beta}(\alpha,\beta)} \\ &= \frac{\text{Beta}(\alpha+k, \beta+n-k)}{\text{Beta}(\alpha,\beta)} \end{aligned}$$

which is a generalization of the derivation for a uniform prior.

Appendix B
Construction of Receiver Operating Characteristic (ROC) curves

The exact procedure for constructing the ROC curves was as follows. Examining the distribution of LLRs in Figure 1 (c), a decision-maker needs to use the LLR of a data set to arrive at a decision about whether the data were generated by a random or systematic process. For example, one approach would be to use zero as a threshold and judge any data set with a positive LLR as being systematically generated and any data set with a negative LLR as being randomly generated. This strategy is equivalent to applying Bayes' rule, as in Equation 1, and choosing the hypothesis with highest posterior probability, assuming the prior probabilities of the two processes are equal. This strategy would lead the decision-maker to correctly classify those systematically generated data sets with positive LLRs (termed a *hit*) but incorrectly classify those randomly generated data sets that happen to have positive LLRs (termed a *false alarm*). Comparing the *proportion* of correct identifications of structure (the hit rate) to the proportion of inaccurate inferences of structure from randomly generated data (the false alarm rate) indicates how good discrimination of random and systematic processes is. Each point on the ROC curve is a plot of the hit rate against the false alarm rate for one threshold on the LLR (in this case the thresholds range from -20 to +20), giving a broad picture of the ability to make accurate judgments about which process underlies observed data.

Appendix C
Modeling the alternation bias

Model of the alternation bias in Experiment 3A

Data from Experiment 2 of Falk and Konold (1997) were used to fit the α and β parameters to model the magnitude of the human alternation bias. Figure 7 shows participants' ratings of apparent randomness (AR) for sequences with varying numbers of alternations, along with predictions of the uniform and biased model. Model predictions on Figure 7 are LLRs which were scaled to the same range as human ratings. The uniform model does not incorporate a bias to judge repetitions as systematic (h_1: $p \sim \text{Uniform}(0,1)$, where p denotes $P(\text{repetition})$) and so produces LLRs that show a similar pattern to that of Falk and Konold's data, but does not capture the human bias to judge alternations as more random. (apparent in Figure 7).

The selection of α and β to model the alternation bias could be done in two ways: selecting values that minimize the average squared difference between scaled LLRs and AR ratings, or selecting values that maximized the correlation between scaled LLRs and AR ratings. Both approaches produced equivalent results: the same range of α and β values minimized squared error and maximized correlation. In this range of values β was approximately one and a half times larger than α. We chose $\alpha = 10$ and $\beta = 15$ as intermediate values in this range and used it in the model for Experiment 3A and 3B. Figure 7 shows that the scaled LLRs for this *biased* model show the same alternation bias as the AR ratings. In summary, the *biased* model of people's biased inferences about randomness versus dependence (as in Experiment 2) represented the nested hypotheses as h_0: $p = 0.5$ and h_1: $p \sim \text{Beta}(10,15)$.

Model of the alternation bias in Experiment 3B

The parameters α and β were selected to accurately reflect participants' alternation bias in the specific task used in Experiment 3A. The data from the nested condition in Experiment 2 were used to choose values of α and β that aligned the model's posterior probability of a sequence being systematic with the proportion of people who identified the sequence as non-random. For each sequence, the number of people who judged it as non-random was assumed to follow a binomial distribution, where the probability of a "success" was the model's posterior probability that the sequence was non-random. The values of α and β identified were those that set the model's posterior probability to maximize the likelihood of people's actual responses. The values obtained were $\alpha = 3.1$ and $\beta = 4.8$.