Utrecht University

Graduate School of Natural Sciences

Exploring Sampling Algorithms to explain Cognitive Characteristics in Random Number Sequences and a Time-estimation Task

Victoria Eshelby (6371833)

MSc. Artificial Intelligence

Supervisors:

Dr. Adam Sanborn
University of Warwick
Dr. Leendert van Maanen
Utrecht University
Dr. Chris Janssen
Utrecht University

July 4th, 2020

Contents

0.1 Abstract

Recently, sampling algorithms have demonstrated their ability in simulating ubiquitous cognitive characteristics, such as autocorrelations and levy-like distributions, found in foraging behaviours. However, this effect has been observed independently in individual tasks. Exploring the co-occurrence of such phenomena has yet to be investigated. This paper explores three main questions: Whether two foraging characteristics co-occur, whether cognitive load impacts said foraging characteristics and to what extent can sampling algorithms explain these cognitive characteristics. Seven participants were given two sequential tasks (Random number generation (RNGT) and a metronome tapping task) separately as well as together. The findings suggest that foraging does not co-occur; that autocorrelations are present in the tapping task but not RNGT and heavy tailed distributions are present in RNGT but not in the tapping task. Cognitive load only plays a role in the tapping task. Further analysis explored task-specific cognitive characteristics outlining potential strategies and patterns participants used (including randomness, run length, pattern type and jump length) and the findings were compared to four sampling algorithms: A Direct Sampler (DS), Markov Chain Monte Carlo (MCMC), Metropolis-Coupled Markov Chain Monte Carlo (MC^3) and a No-U turn Hamiltonian Monte Carlo (HMC). The DS, MC^3 and HMC were able to produce similar results to human behaviour but identifies that a hybrid approach of these three samplers might simulate the metrics produced by humans better.

1 Introduction

Sampling related algorithms have been successful in solving complex problems regarding artificial intelligence (Lai and Spanier, 2000; Yuan and Druzdzel, 2006). Sampling is the process of approximating an expected output given a limited number of samples when provided a posterior distribution. In cognitive science, research has also demonstrated the capability of sampling algorithms as an explanation to foraging behaviour (Van De Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli, 2017; Wagenmakers, Farrell, and Ratcliff, 2004). Specifically, it is believed that sampling may provide a feasible approach to the underlying processes and mechanisms of how the brain generates and selects decisions (Zhu, Sanborn, and Chater, 2018b). Research into this area is promising but the theory is still in its infancy. As such, there are largely unresolved questions in the sampling framework (see Sanborn & Chater 2016 for more details) (A. N. Sanborn and Chater, 2016). Addressing these issues will naturally extend the sampling framework within cognitive modelling and thus generate more human-like output. Currently the focus of sampling algorithms has worked on reproducing two mental foraging phenomena which have been explored in a variety of cognitive tasks (Gilden, Thornton, and Mallon, 1995; Kello et al., 2010; Wagenmakers et al., 2004). These phenomena examine the exploration and search space over a time series, otherwise referred to as levy-flight patterns, in mental representations (*"How far do I travel in a mental space?"*)) and the relationship between different checkpoints, otherwise referred to as autocorrelations, within the mental space (*"How likely am I to return from Location A given I am at Location D at time point I?"*)). Zhu et al., (2018) were able to identify that these two cognitive characteristics are central to identifying processing behaviour albeit they identified that sampling was able to reproduce the characteristics independently in different tasks. As these two phenomena are situated within foraging, it should also be expected that these effects also co-occur. However, to the knowledge of the researcher, this has yet to be explored. Thus to answer this, this thesis extends the previous research by introducing a dual task component.

The metrics used to describe foraging behaviour provide a diagnostic perspective into the underlying cognitive processes, but they do not identify features that are task-specific. For instance, in random number generation tasks (RNGT), sequences generated by human participants, are susceptible to variations of different biases resulting in non-mathematical random outputs (Cooper, 2016; N. Towse and Valentine, 1997). Psychological research has used this task to explore working memory and attention systems (Baddeley, 1966; Daniels, Witt, Wolff, Jansen, and Deuschl, 2003). Cognitive models try to explain these effects by aiming to predict and replicate sequences using pattern matching techniques to varying success (Cooper, 2016; Loetscher, Schwarz, Schubiger, and Brugger, 2008). For instance, the Damerau-Levenshtein model, trained on a sequence length of 300 numbers, predicted up to 11-45% of the next item in a sequence depending on the history length (Marc-André Schulz, Schmalbach, Brugger, and Witt, 2012). A score over 11% was the above chance of randomly predicting the value. However, these models do not attempt to explain the cause of these biases beyond the basic principle that the brain follows selective heuristics and rules when generating sequences and expressions (Jahanshahi et al., 1998; Maehara, Saito, and Towse, 2019; Marc-André Schulz et al., 2012). Sampling algorithms could serve as an explanation to how these biases occur (A. N. Sanborn and Chater, 2016). Whilst there are many algorithms that encompass a sampling model, not all of these approaches are representative of human-sampling behaviour. Zhu et al. (2018) identified that the Metropolis-coupled Markov Chain Monte-Carlo (MC^3), a type of sampling algorithm, was capable of reproducing certain foraging characteristics whereas other sampling algorithms failed to reproduce these effects. This is important as research have approached sampling predominantly using Markov Chain Monte-Carlo (MCMC) as the selected algorithm. Thus it is important to determine what algorithm works best and when it is capable of explaining cognitive processes (Shi, Griffiths, Feldman, and Sanborn, 2010; Zhu, Sanborn, and Chater, 2018a).

For this paper, I arbitrarily take two experimental tasks (a timing estimation task and a generation task) from the Zhu et al. (2018) paper. For the justification behind these specific tasks, see appendix 7.1). This research takes four sampling algorithms from the literature and compares the responses to the data generated from the experimental tasks. In doing so, it addresses three main questions:

- Is there a co-occurrence in foraging behaviours (autocorrelations and search paths)?

- Does cognitive load (using a capacity sharing model) impact the performance of cognitive characteristics?
- To what extent can sampling algorithms explain cognitive characteristics?

The outline of this thesis is as follows. The introduction provides an overview of sampling within a Bayesian perspective, an overview of the sampling algorithms selected, and an in-depth review of the cognitive characteristics (and features). Succeeding sections introduce the experiment, separating the results into the human results then sampling, followed by the discussion of the findings.

1.1 Sampling in the scope of Bayesian Analysis

Humans do not act "fully" Bayesian. This means that humans do not always follow rational principles (Gigerenzer, 2006). A recent paper by Sanborn and Chater (2016) recognised that the past two to three decades in cognitive sciences has seen an explosion of Bayesian literature able to successfully capture and model human behaviour (A. N. Sanborn and Chater, 2016; Van De Schoot et al., 2017). In decision making, this is of particular interest because human probabilistic models, that explain reasoning, has simultaneously been seen as extremely promising and extremely limiting. It is promising because of its contribution in a wide variety of domains including perception, memory, categorisation, reasoning and physics where it has successfully demonstrated compatibility in predicting and simulating human behaviour (Battaglia, Hamrick, and Tenenbaum, 2013; Griffiths, Steyvers, and Tenenbaum, 2007; Wulff, Hills, and Hertwig, 2020). It is limited in that there are still prominent loose ends to its explanation. For instance, significant evidence over the last 50 years have shown that humans do not always follow the rational norms, which probabilistic models would expect (Kahneman and Tversky, 2013). If human behaviour can be captured using probabilistic approaches through Bayesian models then how are we also prone to making systematic errors?

Sampling addresses the inaccuracies created by heuristics and outcomes of incomplete information using finite samples. Finite sampling means that there will be a limited number of samples calculated from a posterior distribution (or hypothesis space) (Zhu et al., 2018a). But rather than calculate and update each individual hypothesis through explicit computations, as performed in Probabilistic inference, a sampling approach approximates over a sampled posterior distribution. Lets explore sampling through a hypothetical example. If one were to estimate the likelihood of a bus being late, they may take a few examples of when a bus is late or they may take many examples. Given that one example represents a single sample, selectively, one creates a list (of samples with a given length) understanding that it is not the true value but an approximation of it. The true value is only created if one samples over an infinite list and realistically, it is impossible to generate an infinite list (without expending an infinite amount of resources to answer it). So the length of this list is constrained by a number of factors such as time and mental resources (Lieder and Griffiths, 2020). On the one hand, it means that a larger list will usually prove more reliable than a shorter list but will also use up more costly cognitive resources. On the other hand, the information from the list will always be incomplete and therefore imperfect. Decision making and decision generation is thus, rationally bounded (A. N. Sanborn, Griffiths, and Navarro, 2010). As explored in the illustration, there are a number of trade-offs which become dependent on the factors of a situation/task. These factors influence the length of a list, and the decisions needed to be made in the aftermath of brainstorming (Gershman, Horvitz, and Tenenbaum, 2015; Vul, Goodman, Griffiths, and Tenenbaum, 2014).

Using a finite number of samples can also explain the trade-off. Vul et al., (2014) identified that humans were more likely to make quick sub-optimal decisions based on fewer samples as the global optimal strategy over long periods of times. Fewer samples are less costly but can increase the chance of generating misleading outcomes and/or biases. Biases can also arise from the starting point of sampling (Lieder, Griffiths, and Goodman, 2012). For example, returning to the bus example, if one encounters an increase in tardiness of their waiting times for a bus over a shorter period of time (or bases their examples on recency and other dominant heuristics), their experiences will determine the starting point and consequently the type of samples. Thus in a sampler the same is also said that each posterior distribution will bias the starting point - for better or for worse. As it explores an n-dimensional space, if it lands on high probability values, then its output will represent a more representative value. Conversely, if it lands on low probability values,

then its output will represent a less successful hypothesis and/or an increase in inaccuracies. In nature, the exploration (search) process is considered sequential (Freidin, Aw, and Kacelnik, 2009) and the movement of exploration consequently shows repetitions based on the starting value which represent long-range or short-range autocorrelations (Ward and Greenwood, 2007). Each approximation account for the inconsistencies observed in many decision making journals and provides empirical evidence to human-bounded rationality (Gigerenzer and Goldstein, 1999; Parpart, Jones, and Love, 2018; Saposnik, Redelmeier, Ruff, and Tobler, 2016).

There is something stochastic in human nature that mismatches pure probabilistic theory. Its reputation, however, indicates that to some extent, humans are able to perform *most*) inductive computations similar to a probabilistic framework (Tenenbaum, Kemp, Griffiths, and Goodman, 2011) albeit noisily. In a systematic review evaluating Bayesian literature, Van de Schoot et al. (2017) identified increased trends ranging from Bayesian approximation models to successful simulation using participant data to model behaviour. A Bayesian account for human decision making can no longer be taken at its literal computation for that would entail concrete complex calculations (using Bayes theorem) to be performed for all possible outcomes given a scenario. Van de Schoot et al., analysis' details the shift in cognitive models using absolute probability to one of approximation. The rise in approximation frameworks has enabled a plethora of research dedicated to heuristics and mistakes in judgement tasks to be incorporated within Bayesian models; answering the paradoxical debate that once existed between probabilistic frameworks and heuristics (Kahneman and Tversky, 2013). By using approximations instead of explicit calculations, one is able to handle increasing complexity because the computational cost, of these approximations, scales with the number of samples rather than the size of the posterior distribution. Thus using approximations means that generating large computations no longer seems infeasible. Bayesian modeling can therefore become scalable albeit more error prone. More succinctly, sampling deviates from an ideal Bayesian model as it approximates over a sampled posterior distribution rather than through explicit symbolic calculations. To not be fully Bayesian assumes a Bayesian brain that does not explicitly represent or calculate probabilities but instead through sampling. Sampling takes an approximation of samples from a distribution taken from memory or imagination.

1.2 Sampling algorithms

In total four sampling algorithms have been selected for this research. A direct Sampler (DS), Metropolis-coupled Markov Chain Monte-Carlo (MC^3), a Markov Chain Monte-Carlo (MCMC) and a Hamiltonian Monte-Carlo (HMC). The goal of this section is to provide an intuition regarding the difference between algorithms[1].

A direct sampler is defined by independently drawing samples given the posterior probability distribution. It is a highly efficient algorithm but scales exponentially dependent on the dimensionality of the hypothesis space. This approach has already been explored in probability matching and mental sampling (Vul et al., 2014; Zhu et al., 2018a). An MCMC algorithm samples from its posterior distribution and calculates an expected value. The sampler samples locally, selecting its next value based on whether the sample has a higher probability value than its current position. Dasgupta et al., (2017) identified multiple characteristics in hypothesis sampling and the MCMC. Overall, the types of frequencies of samples that are generated are proportional to the posterior probabilities; that samples are compared between the relative probabilities between two hypothesis and that the output generates autocorrelated samples (Dasgupta, Schulz, and Gershman, 2017). The algorithm moves in locality favouring nearby locations over long range locations. This sampling algorithm has been explored in human decision making, creativity, mindfulness (Kee, Chaturvedi, Wang, and Chen, 2013; A. Sanborn and Griffiths, 2008; Vul et al., 2014), anchoring (Lieder et al., 2012) and memory (Annis, Lenes, Westfall, Criss, and Malmberg, 2015). The MC^3, or parallel tempering involves running multiple Markov chains in parallel usually at different temperatures. Using temperatures are separated between hot and cold temperatures. Hot temperatures allows chains to make larger jumps and cold chains make local steps in the current probability peak. As the MC^3 contains multiple chains, the

[1]For the algorithmic implementations: MC^3: can be found in Zhu et al., 2018 paper; HMC is detailed using NUTs by Hoffman et al., (Hoffman and Gelman, 2014

algorithm incorporates swapping between hot (far reaching jumps) and cold chains (local jumps) between two randomly selected chains to explore the environment better. The swap is either accepted or rejected according to the Metropolis principals. The MC^3 has been explored in task complexity and goal specificity (Arminger and Muthén, 1998) and multi-modal distributions (Zhu et al., 2018a). The MC^3 approach has been considered highly inefficient in exploration and alternative algorithms such as the HMC have worked to improve how to explore the search space. The HMC works by applying momentum to each generated sample. The HMC is a variant of the MC^3, where it is believed to enable more efficient MCMC sampling. The HMC does this by taking the gradient of a probability distribution and uses to propose future states within a Markov chain. The gradient is calculated through approximated Hamiltonian dynamics (which is incorporated through a leap-frog mechanism) (Betancourt, 2017). In other words, the HMC generates new samples by giving its current location momentum. Areas with high modal distributions are provided higher potential energy but smaller kinetic energy and vice versa. This allows it to travel more efficiently in higher dimensions(Chevallier, Pion, and Cazals, 2018). Its momentum starts to change when there is a change in a gradient of a distribution(e.g. Gaussian) which causes the momentum to change. The best way to describe a HMC is a ball moving on a frictionless surface. Consider a Gaussian distribution simulating this valley. As the ball traverses along the slope, its gradient determines how big (or small) of a push it receives. However, when applied in a domain of bounded constraints and when its posterior distribution has a relatively flat gradient, the algorithms efficiency in exploring a domain becomes severely impacted. For instance, in the random number generation task, whilst there are edge cases, there is no hill to slow it down or change in the momentum so the momentum keeps pushing it towards a specific direction skewing the overall results. Multiple algorithms have since been developed to handle situations where there is a constrained set and a uniform distribution such as the NUTS (No U-turn sampler) algorithm which transforms constrained spaces into an unconstrained space and automatically adapts the trajectory length. (Chevallier et al., 2018; Hoffman and Gelman, 2014).

1.3 Cognitive Characteristics in sampling

This section explores two different types of cognitive characteristics. The first explores ubiquitous cognitive characteristics found in foraging. This is where an effect has been found present in a variety of cognitive tasks and can therefore identify underlying mechanisms or processes. The second explores task-specific cognitive characteristics and introduces the two tasks chosen for this experiment (a random number generation task and a tapping task). This identifies cognitive elements that provide insight into the output of a specific task and therefore explores potential features that can differentiate the models between each other and more importantly, the human data.

1.3.1 Ubiquitous Cognitive Characteristics

The focus of this thesis explores two foraging phenomena, sampling has been able to replicate, given a time-series task. These two phenomena are:

- The exploration of the search space.
- The relationship between generated items.

Exploration of a search space

The exploration and foraging of mental patterns has been identified to share a remarkable resemblance between foraging behaviours found in nature(Brabazon, McGarraghy, et al., 2018; Patten, Greer, Likens, Amazeen, and Amazeen, 2020). Researchers believe that animals (and humans alike) have adapted to use efficient search strategies that explore movement especially when resources are patchy and sparsely distributed (Montez, Thompson, and Kello, 2015). Montez et al., 2015 describes efficiency as maximising resources per unit time. Pattern et al. (2020) explored how humans forage in their mental landscape in a free-recall generation task using countries and animal naming tasks. Free-recall tasks matched a heavy-tailed, levy flight like distribution. The authors also identified that the distances between the cosine similarities and or distances in terms of location also matched a heavy tailed distribution (Patten et al., 2020). Viswanathan et al., (1999) has identified Lévy flight as the optimal movement pattern in foraging behaviour and this proposal is generally accepted due to its robustness and ubiquitous presence in many foraging studies (Barthelemy,

Bertolotti, and Wiersma, 2008; Hills, Jones, and Todd, 2012; Montez et al., 2015; Rhee et al., 2011). Within internal foraging this is believed to reduce processing and retrieval time (Humphries, Weimerskirch, and Sims, 2013). The levy flight foraging hypothesis follows a levy distribution. A levy flight is the probability of performing a jump of length l (see equation 1) and levy distributions generate heavy tailed distributions meaning high probabilities are followed by low probabilities.

$$P(l) \sim \frac{1}{l^{\mu}} \tag{1}$$

We apply levy distributions as to how a sampling algorithm explores an mental landscape. As a sampler bases its hypotheses on a posterior distribution, it acts as an excellent starting point in describing the exploration/foraging behaviour a mental landscape undergoes. Sanborn & Chater (2016) described this posterior probability distribution as a mountainous, high-dimensional landscape. Each 'mountain-range' can represent a dimension and each mountain, a high probability distribution. The goal of a sampler is to reach the peak of the mountain range. Thus the highest probability value represents the global peak of a mountain range. A high value also indicates the chance of a more successful hypothesis. A low probability is the space between two peaks that can be far apart or close together. A sampler (or agent) blindly navigates the mountain range to find peaks. The sampler has no memory of where it has been and explores its environment step by step making it terribly inefficient when sampling with replacement[2]. At each step, it lands on a probability value. Each step represents a sample. If this value is higher than its last, then it moves to that location; if it is lower, it returns to its initial value and tries again. This enables a sampler to assess whether it is able to find a higher peak based on the relative probability values. But of course, there are limitations to this approach. A sampler can become convinced it has found the peak of a mountain range when in fact it has only found a local peak of a much smaller mountain; This indeed adds to the rational response and provides a feasible explanation to how biases such as anchoring (Lieder et al., 2012), reasoning fallacies(Dasgupta et al., 2017) and probability matching (Vul et al., 2014) can occur. In other words, the success and errors created by a sampler are dependent on where a sampler starts. It is also dependent on the distance between mountains, the gradient of these mountains, as well as the distance between mountain ranges. In cognition, it is unlikely that these mountain ranges or/and mountains are equally spaced but instead grouped together and then separated by vast valleys of low probability similar to foraging in nature. Animal foraging explores the movement in space and how one traverses the jumps, not only between each successive mountain, but also between each successive mountain range. The jumps between the mountains and mountain ranges therefore should also follow these powerlaw distributions. If a sampling model is able to capture the jumps between patchy distributions, then it stands in good stead that it is also able to serve as an explanation in how cognitive processes move between resources.

The relationship between generated items

The second foraging trait found in many cognitive tasks is the role of relationship between generated items. When data is found in a time-series, this is best explored using autocorrelations. Autocorrelations describe the relationship between sequential samples. More formally, autocorrelations are associated to long-distance serial correlations between values that are k time period apart. In an experiment, this means there are relationships not only after a trial and the n + 1 trial but potentially after n + 500 trials. If we return to the mountain-range illustration, we could perceive this as returning to the same mountains in a mountain range or jumping between mountain ranges previously visited (perhaps because the distances are closer and therefore require less resources under computational constraints). Naturally, it is not always possible to immediately arrive at a mountain peak. There are also many ways up this landscape so the steps to get an agent to the peak are considered noisy variables. Autocorrelations have been observed in a variety of cognitive tasks and tend to follow a 1/f scaling law (Gilden, 1997; Kello et al., 2010; Wagenmakers et al., 2004). Predominantly, this area has been investigated in temporal, spatial and memory related domains. Kello described this law as the ability to define the intricate regularities and dependencies that span multiple temporal and spatial scales using a finger tapping task. If one observes the error

[2]Of course, one can also place further emphasis on various parameters such as the type of jump between steps, certain satisficing acceptance criteria, whether there are multiple agents exploring the same (or different mountain range) and momentum with each step and we explore this using various sampling algorithms (A. N. Sanborn, 2017; Shi et al., 2010). For now, our agent resembles a sampling algorithm with similar movement to a Monte-Carlo Markov Model.

generated from each tap. The deviation in error simulates a normal distribution but if we consider all possible taps, there's clearly non-random behaviours; the sequence of actions no longer represent chance but deterministic behaviour whose features can be evaluated using power spectra. Using a power spectra then enables one to evaluate the overall output of different mechanisms at different scales. In other words, the power spectra acts as a suggestive common principle that underpins the different mechanisms relationship. The ability to move through different mechanisms becomes important because if it acts as a core principal, then being able to replicate this effect will assist in simulating and understanding cognitive processes further.

The lag between the samples correlation can range in a combination of short distance (the time period, k, is small) or long distance (the time period, k, is large) autocorrelations. This dependence usually decays over time. More intuitively, 1/f noise is a spectral powerlaw used to explain the correlation value that occurs between no correlation (white noise) and a random walk (Brownian motion) (Ward and Greenwood, 2007). Equation 2 explains the autocorrelation in a lag-based setting. Using autocorrelations important because the alpha value provides an interpretation of the amount of noise that is distributed over time.

$$C(k) \sim \frac{1}{k^{\alpha}} \tag{2}$$

Where C(k) is the autocorrelation function of temporal lag K. The same phenomenon can also be expressed using frequency:

$$S(f) \sim \frac{1}{f^{\alpha}} \tag{3}$$

Where S(f) is the autocorrelation function of spectral power resulting from a Fourier analysis. The latter (equation 3) is more popular for fitting alpha values. What can be interpreted from this function is the level of noise that occurs which is usually a value between [0,-2]. A value nearer to 0 indicates white noise, that is stochastic behaviour occurring and a value moving towards ∞ identifies short-ranged correlations (only recent trials can impact present behaviour). In between this range ($\alpha \in [-0.5, -1.5]$), generates long range autocorrelations where trial-n contain a non-negligible impact on the present behaviour). If values are higher than 0, then this indicates anti-correlation like behaviour. Anti-correlatory behaviour is a serial negative relationship. This has been shown in nature where a heart-beat have generated long-range anti-correlations, that is large heart beats are compensated by small heartbeats and vice versa(Peng et al., 1993). When participants had heart conditions, often these correlations disappeared and the heart beat intervals remained within the bounds of a levy distribution. A more succinct explanation of: noise, its values, a potential sampling algorithm candidate and the interpretation one can infer from it can be found in Table 1).

Noise	Range of α	Sampling Effect	Explanation	Interpretation
Anti-correlated Noise	$[1, 0]$	Direct-Sampling	Behaviours are similar to anti-correlated noise	Behaviours have a long-ranged negative relationships.
White Noise	$[0, -0.5]$	Direct-Sampling /stochastic	No Autocorrelations	Behaviours are independent to one another
Pink Noise	$[-0.5, -1.5]$	MC^3	Autocorrelations similar to 1/f noise	Behaviours have serial long-range relationships (n = n + 500)
Brownian Noise	$[-1.5, -2.5]$	Random-walk Metropolis	Autocorrelations similar to $1/f^2$ scaling	Behaviours have serial short range relationships (n = n + 3)

Table 1: explores the relationship between autocorrelations, the noise pattern they generate and the interpretation you can infer if you end up with a particular result.

If a sampling model is able to capture autocorrelations, then it stands in good stead that it is also able to serve as an explanation behind the cognitive process in the types of dependent behaviours we unconsciously

explore. In sampling producing similar correlatory patterns has been demonstrated over multi-modal distributions (Zhu et al., 2018a). What can be identified is that sampling is capable of generating cognitive characteristics albeit the scope of research still requires further exploration which this research aims to extend. The point of origin of this research is to address the replication of levy distribution and 1/f noise in cognition using a sampling framework. We extend Zhu et al., research by introducing a secondary task (tapping task) to a generational task. The goal is to evaluate whether these two effects can co-occur in multiple tasks and one way to evaluate this is by adding dual tasking. By adding an additional element of complexity, not only are we able to test whether there is a difference in sampling behaviours but also if these effects are simultaneously present, we can further contribute to whether mental sampling is sequential or parallel?

1.3.2 Task-Specific Cognitive Characteristics

On one hand, exploring ubiquitous characteristics may provide an overall insight when comparing performances of a model between different tasks. On the other hand, it does not explore the features generated in the task. Exploring task-specific characteristics can add an additional metric into evaluating how a computational model performs as well as identify specific strategies a participant might incorporate(Alves, Tassini, Aedo-Jury, and Bueno, 2020). We return to the mountain range analogy. The ubiquitous characteristics tell us how we have travelled between mountains but as mentioned there are noisy variables at play. The task specific characteristics can be treated as the path travelling up a mountain. One could suspect, there are shortcuts or a path that are more easily traversed along a mountain to reach the peak and this provides additional information that can be investigated more thoroughly. For instance some patterns in a RNGT are unique to a participant(Marc-Andre Schulz, Baier, Böhme, Bzdok, and Witt, 2020; Marc-André Schulz et al., 2012). Schultz et al. (2012) likened person-specific patterns to a bio-metric feature where an n-gram Monte-Carlo was able to predict some participants independent sequences better than others where exploring sequences in participants demonstrated the importance of identifying patterns of up to 88% accuracy. Random number generation tasks have been evaluated through a variety of methodology and in humans, random number generation has been shown be susceptible to a variety of biases (Cooper, 2016; Figurska, Stańczyk, and Kulesza, 2008; N. Towse and Valentine, 1997; Marc-André Schulz et al., 2012).

Towse & Neil (1998), hereunto referred to as T&N, compiled a review where many of the above cases of bias have been determined through a variety of these metrics. Here we explore redundancy (R), Random number generation for bigrams (RNG), the Turning Point Index (TPI) and Adjacency (A). Towse focused on a RNGT that set ranged from 1-10. The R score from the T&N paper ranges from 0 to 100 % where 0 represent equal frequency (complete randomness on an infinite scale) and 100% represents responses that are predictable/identical. R will always be slightly bigger than 0 on a finite sequence length and this value moves towards 0 as the sequence reaches infinity. Evans (1978) explained RNG as an alternative to R but instead investigated the change in bigrams (such as the occurrence of the number 1,2) by the change in sets between response (occurrence of the number 2). The RNG score ranges from 0 (completely random scores) to 1(completely predictable scores). Both R and RNG are designed to evaluate deviations between the randomness of response and bigram responses. Alternative approaches to explore RNG is the RQA (recurrence quantification analysis) within time series which has been demonstrated to describe executive functions in alignment to more traditional measures however has yet to be thoroughly reviewed (Oomens, Maes, Hasselman, and Egger, 2015). Adjacency evaluates all possible response pairing to evaluate bigram biases where 0 identifies no neighbouring pairs and 1 if the sequence is composed entirely of pairs. TPI measures the frequency of a response changing (either ascending and descending) within a series. A score of 1 indicates an unbiased sequence where the actual output matches the expected number of switches among a chain. A value less than 1 indicates that there are fewer turning points than expected where participants are more likely to make more runs than the expected amount (the sequence 1,2,3,4,5,6,2 has 1 turning point -at position 6 [6]- and a TPI score of .3). A value higher than 1 means there are more turning points (the sequence 1,8,3,6,8,7 has 3 turning points - at positions 2 [8],3 [3], and 5 [8] - and a score of 1.25).

The T&N paper hints, to some extent, a sort of momentum however evaluating higher order distribu-

tions remained outside the scope. Multiple studies have since identified that random number generation in humans is biased to a certain degree (Baddeley, 1966; Cooper, 2016). Human perception of randomness is skewed(Hahn and Warren, 2009). For instance, participants were less likely to repeat a response than expected in an actual random sequence (inhibition of return as found in memory tasks (Johnson et al., 2013)), more likely to transition between neighbouring pairs and that bias is more likely to happen in larger sets than smaller sets. Also, what confirms the concept of momentum is that one is less likely to switch their actions when producing sequential numbers and in dual tasking, this effect is amplified (Cooper, 2016). If this momentum exists, to what extent does it exist? For instance, in examining the frequency of runs length as questioned by Schulz et al. (2012) or type of preference in a pattern or preference in the jumps of patterns?

Depending on the type of sampling algorithms will change the way in how it interprets a number. Direct sampling produces a truer random sequence. A MC^3, HMC and MCMC will produce dependent samples. The HMC,MCMC and MC^3 are capable of the biases between neighbouring pairs. In the MCMC, the chain is always moving along a distribution is based on adjacent values neighbours. For the HMC and MC^3, the movement across a distribution will be based on the number of chains in use and for the HMC, the energy added. Within a MC^3 given that it's basing its distribution on a uniform distribution (as the probability from jumping between 1-10 will always be 0.11, or 11% for each element), the proposal will always be accepted under the assumption the proposal is within the boundaries specified. This can explain the self-avoidance of repeating numbers as each new proposal is accepted. In a non-traditional HMC, this acceptance rate is changed due to transforming a uniform distribution to handle bounded environments. When evaluating features, the run length can be used to distinguish between a metropolis and HMC as the lengths indicate momentum through sequences. Thus one would assume there to be more runs in the HMC than a MC^3, DS or MCMC. The same reasoning for distinguishing MC^3 to HMC could be used through the pattern types. Instead of assessing momentum, one could also use pattern types to identify the probability of moving from the current state to the proposed state. In the MC^3 the probability for jumping between the sequence 1,2,1 equally as likely as the sequence 1,2,3 or 1,2,8, whereas the HMC would output the ascending or descending patterns as being more likely. To explore the differences between MC^3 and MCMC, we explore the jump ranges in the adjacency values. The MC^3 would also treat the jump sizes 1,2 vs 1,8 equally likely, whereas a MCMC would favour shorter jumps.

2 Experimental Methodology

This section evaluates the approach on collecting human data only. In total 7 participants took part in the experiment and were recruited using convenience sampling.

2.1 Materials and Data Collection

The experiment was designed using psychopy and ran within the psychopy standalone environment (Peirce, 2007). The data collected included a timestamp (in seconds) of when the participant pressed the spacebar, the duration of the spacebar press and when the participant released it. The experiment consisted of three tasks which each ran for 394 seconds. Two tasks are grouped under a single/control condition and one task was grouped as a dual condition. Voice capture using a microphone was enabled to record their voice. The data collected included the order, a time-stamp for each space bar pressed and released, interkeypress interval/ inter-response intervals (IRI) between each key press. Audio data, which was an audio file, was collected for the length of the task time, was used to transcribe each number generation. The transcription process matched audio to the key press times dependent on when a participant spoke a number. The experiment was completed under 'lab conditions' [3].

2.2 Procedure

The order of the tasks were randomised to counterbalance any order effects.

[3]The experiment had been set up for lab conditions and whilst this was specified in the introductory form, participants ran the code on their home devices due to external conditions.

Random Number Generation. In the RNG task, participants heard a computer-generated beep every 750 ms for the duration of the task. This was to assist them in maintaining consistency and rhythm. Their goal was to say aloud a number ranging from 1 to 10 for each beep such that the string of numbers would be in as random an order as possible. This experiment used Baddeley's (1966) initial instructions, which ensures the concept of randomness (with replacement): Participants were given the analogy of picking a number out of a hat, reading it out loud, putting it back, and then picking another. This is the first single task.

Time-interval estimation. In the tapping task, participants received a brief practice period consisting of 30 beeps at an interval of 750 ms to assist them in maintaining consistency and rhythm. Their task was to press the spacebar to reproduce this interval without feedback. Visual feedback (in the form of a circle appearing on the screen) was provided to acknowledge each key press. This is the second single task.

Random Number generation + Time-Interval estimation. In the RNG + Tapping task, participants were required to simultaneously perform the Random Number Generation and Time-interval estimation. Participants received a brief practice period consisting of 30 beeps at the start of 750 ms to assist them in maintaining consistence and rhythm. Participants were expected to press the spacebar and say aloud a number ranging from 1-10. Visual feedback (in the form of a circle) was provided to acknowledge each key press. This is the dual task condition.

2.3 Data Pre-processing and metrics

Data Preprocessing. After data collection, only valid elements were included. Valid elements included sequentially recalling numbers within the range of 1-10. Elements that did not meet this criteria were removed and the interval time was added to reflect the response time in IRI of the next valid entry. It should be noted that, 6/7 participants performed combined these tasks in parallel (spacebar and "1"). 1/7 participants performed these tasks sequentially (spacebar then "1"). All participants were included in the findings.

Metrics. The metrics used in this exploration are the two central cognitive characteristics. Heavy tailed distributions were calculated using the powerlaw package (Alstott, Bullmore, and Plenz, 2014) for model evaluation and fitting necessary parameters. Auto-correlations extracting coefficients and intercepts were modelled after Turvey & Rhodes and provided by the Warwick University, sampling team (Leon, 2020; Rhodes and Turvey, 2007). Finally to explore randomness in the random number generation. Several metrics (R, A, RNG, TPI) were incorporated from Towse & Neils (T&N) and other metrics (phase space and autocorrelation) from Morariu et al. (1995,1999) were used . Effect sizes were evaluated using hedges g unless otherwise specified.

The additional patterns used to investigate differences between sampling algorithms include run length ($n+1, n+2, n+3$ vs. $n+1, n+2, n+3, n+4$), adjacency or bigram jumps(1,2 vs 1,5 vs. 1,8), pattern type (n+1,n+2,n+1 vs. n+1,n+2,n+3). These patterns were identified within the sequence using the Knuth-Morris-Pratt algorithm where further pre-processing was applied to avoid mutual dependencies in the frequency count. For instance, if mutual dependencies were included, the sequence [1,2,3,4,1,2,3,1,2] would output the pattern [1,2] three times instead of reporting once.

3 Experimental Results

3.1 Descriptive Statistics and Preliminary Analyses

The single task number generation task produced response lengths ranging from 441 to 521, with a mean of 488 items. The tapping task (labelled as IRI) produced an average of 591 items ranging from 488 - 682 with a mean of 1 press per every .52 seconds. The dual task produced sequence lengths ranging from 310 to 559, with a mean of 473 items and 1 press for every (on average) .7 seconds. The expected output for all 3 experiments was 512, assuming participants were able to exactly represent a .75 second gap interval for 384 seconds. See table 7 for specific details regarding the descriptive statistics about each participant in

the appendix. As shown in table 7, both response times for the base-line and dual task had IRIs below the expected length. Response times in the dual task IRI were significantly closer to the single task IRI value, t= - 1.783, $p < .05$. The effect size for IRI response time is medium, d = .661. The differences between number item generation as a base-line and dual task was not significant (t = .34, p = .37). This identifies that participants were more likely to use the number generation task as a method to produce a rhythmic interval or focused less on anticipating the next task.

3.2 Cognitive characteristic metrics

Table 2 outlines the overall findings for the cognitive characteristics related to foraging behaviours.

Cognitive Characteristic	**Tapping Task**		**RNG**	
	Single	Dual	Single	Dual
Powerlaw Distributions	9.49	3.98	3.42*	3.46*
Power-Spectra	-1.04*	-.51*	.23	.3

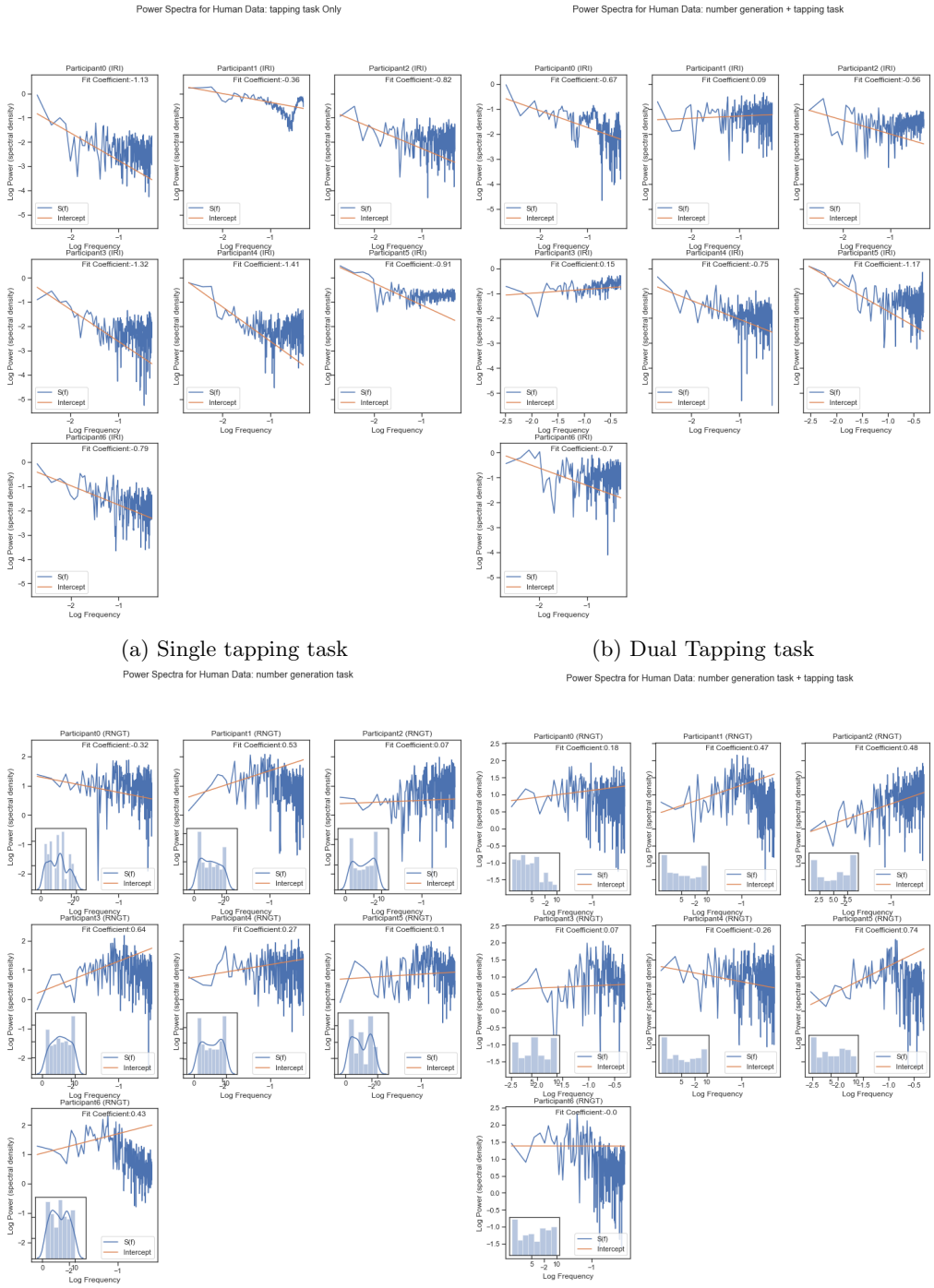
Table 2: Outlines the overall results for the cognitive characteristics autocorrelations and levy flights. Shaded areas indicates values that lie within expected boundaries (Levy: agglomerated μ scores that lie within the range of [0,3] ; Autocorrelations, defined using a Fourier's power-spectral transformations: α scores with expected boundaries between [-3, -.5]). * represent individual participant values that scored within the expected boundaries.

3.2.1 Power-spectra

After confirming that both data sets satisfied the levene's test for homogeneity of variance for the number intercept (F(1,12) = 1.06, p= .324), two separate independent t-tests were run to compare the difference in intercepts between dual tasking and single tasking in number generation to calculate the long-range auto-correlations in the signal. A visual representation of the scores can be seen in Figure 1.

Tapping task vs Number Generation + Tapping Task
In the single condition, the coefficient value ranged between $\alpha \in [-1.41, -.36]$, mean = -1.01. In the dual condition, the coefficient value ranged between $\alpha \in [-1.17, .148]$, mean = -.51. Figure 1a and 1b shows participants results for autocorrelation evaluating only the output from the tapping task data (IRI). This indicates an average on the border of white/pink noise for dual conditions and pink noise in the single condition. A t-test determined that there was a significant effect in the number generation task, t = -1.976, $p < .05$. This indicates that dual/single tasking plays a role in IRI generation output where long range auto-correlations disappear. Using IRI produce long-range autocorrelations in single tasks and become more stochastic (white noise) when a second task is introduced.

Number Generation vs. Number generation + Tapping Task
In the single condition, the coefficient value ranged between $\alpha \in [-.1895, .53]$, mean = .23. In the dual task condition, the coefficients ranged between $\alpha \in [-.25, .73]$, mean = .3. This indicates values in the range of white-noise to long-range serial anti-correlations. Figure 1c and 1d shows participants results for autocorrelation evaluating only the output from the random number generation task. A t-test determined that there was no significant effect in the number generation task, t = .0351, p = .486. This indicates that dual/single tasking does not play a role in random number generation output. From these results, autocorrelations do not occur in random number generation tasks. Morariu's approach to defining autocorrelations yielded values of .15 and .2 for the single and dual task, respectively.



(a) Single tapping task

(b) Dual Tapping task

(c) Single Random Number Generation

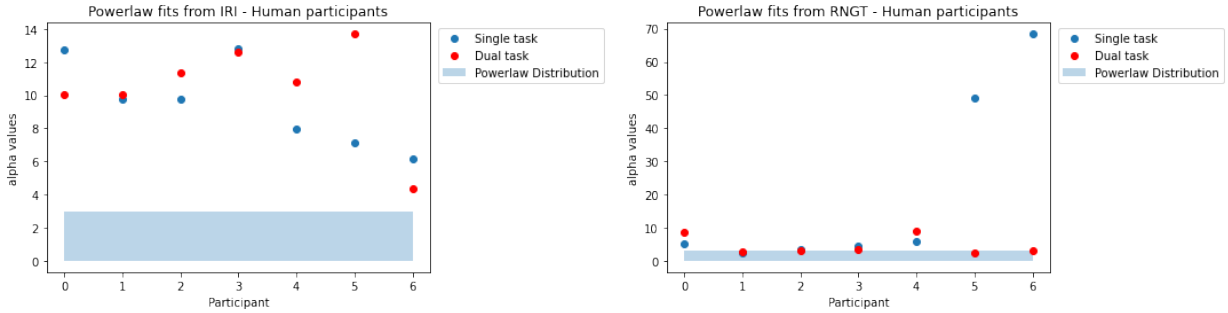
(d) Dual Random Number Generation

Figure 1: TL: IRI + single condition; TR: IRI + dual task condition; BL: Distances + single condition; BR: Distances + dual condition. The top figures represent the IRI values taken from the tapping task section, the bottom figures represent the number values taken from the number generation section. On the left are the single conditions, and on the right is the dual task condition. The graph is represented using a log-log plot. Each plot represents an individual participant's performance of the autocorrelations. The S(f) value represents the frequency over time. The intercept value represents the coefficient output. The bottom two figures contain histogram subplots of the frequency distribution from the set of 1-10 generated by the participant. Coefficient values were either anti-correlated, white noise or pink-noise.

3.2.2 Heavy Tailed distributions

Tapping task vs Number Generation + Tapping Task
In the single condition, the powerlaw exponent ranged between [6.18, 12.79]. The overall aggregated time series had a μ score of 9.5. All individual values scored a higher powerlaw exponent $\mu > 3$. In the dual condition, the powerlaw exponent range between [4.35, 13.7]. The overall aggregated time series had a μ score of 3.98. The powerlaw exponent $\mu > 3$. None of the participants, as shown in figure 2b, scored within the heavy-tailed distribution boundary indicating that the tapping task is unlikely to represent neither a heavy nor a light tailed distribution.

Number Generation vs. Number generation + Tapping Task
The powerlaw exponent was calculated by subtracting the difference between two absolute, consecutive numbers (b-a). In the single condition, the powerlaw exponent ranged between [2.29, 77.4]. The overall aggregated time series had a μ score of 3.42. The powerlaw exponent $\mu > 3$. In the dual condition, the powerlaw exponent range between [2.39, 6.85]. The overall aggregated time series had a μ score of = 3.45. The powerlaw exponent $\mu > 3$. Figure 2a In the single condition, 1/7 participants produced a score that would fit a potential heavy-tailed distributions. A distribution comparison determined that the distribution resembled a powerlaw distribution better than an exponential distribution, R = 3.35, $p < .001$. In the dual condition, 2/7 participants produced powerlaw values also indicating heavy-tailed distributions over exponential distribution (R = 8.25 and 5.2, $p < .0001$).



(a) Tapping task powerlaw values

(b) RNGT powerlaw values

Figure 2: shows the μ values from the Powerlaw analysis for all 7 participants. Each point represents a participants powerlaw fit. Red indicates The blue box indicates whether the value fits inside a powerlaw distribution thus indicating a potential heavy-tailed fit.

3.2.3 Randomness Metrics

This section relates to only the random number generation as extensive research has explored the random section. Overall participants explored the phase-space of the random number generation task fairly well over the conditions producing (on average) 89/100 pairs in the single condition and 91/100 pairs in the dual condition. Table 3 outlines the metrics used to evaluate the randomness. In the single task, the adjacency score ranged between .19 and .48, mean = .35 This indicates that participants were more likely to choose non-neighbouring pairs over neighbouring pairs. The redundancy value score ranged between 0.29% and 5.02%, the mean ranged at 2.4 % which suggests that many of the participants were able to score near equality and therefore the score indicates near randomness. The Turning point index ranged between .61 and .99, mean = .81. This indicates that the actual value was lower than the predicted value suggesting evidence of momentum in participants recall. In the dual task, the adjacency score ranged between .17 and .44, mean = .32 Similar to the single task, this indicates that participants were more likely to choose non-neighbouring pairs over neighbouring pairs. The redundancy value score ranged between .29% and 5.52%, the mean ranged at 2.62%. The Turning point index ranged between .66 and 1.04, mean = .85. This indicates that for 6/7 participants the actual value was lower than the predicted value suggesting momentum in participants recall.

Condition	A	RNG	R	TP	Auto
Single Task	0.345	0.506	2.401	0.814	0.195
Dual Task	0.327	0.487	2.624	0.849	0.155

Table 3: Comparing the two condition of the single task and the dual task to the various metrics. A represents adjacency, RNG, equality of bigram use, R is the redunancy, TPI represents the turning point index from the T&N paper and Auto represents the autocorrelation function from the Morariu et al (1995) paper.

A statistical t-test determined that there was no significant effect in randomness scores between the single or dual condition in R (t = -.12, p = .453), A (t = .584, p = .28), and TPI value (t = -.43, p =.334).

3.2.4 Pattern Metrics

The condition (dual or single task) was measured against three pattern types. A shapiro test was used to evaluate the distribution of the pattern lengths, pattern types and pattern jumps. In all three cases, the distributions did not reflect a normal distribution of residuals (R = .803, $p < .001$ for pattern 1, R =.86, $p < .001$ for pattern 2 and R = .88 , $p < .001$). This is understandable given that previous research identifies that individuals are less likely to repeat pairs of long numbers and this also applies across distributions. Statistically, one is more likely to mention a set of pairs than produce a run of 5 digits. To account for inequality a one-way ancova was ran to assess the run lengths in the patterns between the conditions. What is descriptively interesting is that participants on average were more likely to provide pattern runs of three (88:104, single: dual task respectively) than two (67:64, single:dual task respectively), F(1,109) = 109.53, $p < .001$. However, the overall effect size is 0 indicating very little difference between the run lengths. There is no statistical difference between the two conditions, t(14) = .308, p = .766. There was no statistical difference in the conditions, p=.322 when examining pattern types. Between each participant, the counts of observations varied, where the frequency of pattern type between the two classes shared an uneven distribution (184 for n+1,n+2,n+1 by 454 for n+1,n+2,n+3) as well as an uneven distribution across participants (a range from 46 to 128). A 2 x 2 ancova revealed a statistical effect in the pattern type F(1,635) = 89.67, $p < .001$, with an effect size of .123. This indicates that participants are more likely to ascend or descend when deciding patterns and therefore indicates to some extent, momentum. A 2 x 6 ancova revealed there was no statistical difference between frequency of bigram pairs in the dual or single task, T(14) = 1.078, p = .281 when examining pattern jumps in bigram pairs. There is a statistical difference in the pattern jumps, F(2,7) = 63.79.2, $p < .001$, with an effect size of .06. There were differences in frequencies between short and long jumps (t(1037.4) = 8.514, short and medium jumps (t(100.84) = 5.361) of both medium effect sizes (.531 and .672 respectively) but not between medium and long jumps (t(99.029) =0.387, $p = .9$ Bigrams that are grouped closer (e.g. 1,2) together are more likely to be successively joined than bigrams that have a further distance (e.g 1.8).

3.2.5 Brief Overview

The results identify potential powerlaw distributions that fit within a levy-flight and the power-spectra indicates Anti-correlated noise. There is relatively little difference between the single and dual task except when observing power-spectra results in the tapping task. In the next section, we incorporate only the data taken from the human RNG single task so as to not interfere with the reproducibility of the methodology as a comparison.

4 Sampling Methodology

This section evaluates the result against three sampling algorithms: Direct sampler (DS), MCMC, MC^3 and HMC and incorporates these models in the Random Number Generation task (RNGT). Each model presents output generated as a simulated participant given the sampler. The focus of these algorithms is not to optimise the parameters to simulate the closest value to "human data" but to evaluate whether

they are able to reproduce the effects observed in the experiment based on the heavy-tailed distribution, autocorrelation and randomness. All three models had their targeted distribution set to sample from a uniform distribution and starting location at 1 and produced a sequence of 512. The exception was the direct sampler whose starting location was randomly generated. The MCMC's acceptance criteria ranged between 89-90%, the MC3 whose range was at 99% and the HMC whose range was at approximately 75%. In total 50 models were run for each algorithm type. Unless specified, the means from the 50 averages were used as a comparison.

4.1 Model selection

The MCMC and MC^3 were modelled after Zhu et al., (2018) paper. The MC^3 algorithm deviates from the Zhu et al., (2018) paper by swapping the chain every iteration. The HMC was modelled using the NUTS algorithm where the target distribution acceptance rate was set to 90% and tuning to 1024, the number of chains were set to 2. The direct sampler was generated using the standard random packages python provides.

5 Sampling Results

Table 4 outlines the overall findings for the cognitive characteristics and compares the values found to human data.

	Human	**DS**	**MCMC**	**MC3**	**HMC**
Powerlaw distributions	3.46 *	4.02*	76.48	37.16 *	3.76*
Power Spectra	.24	.00	-1.29	-1.28	-0.22

Table 4: shows the overall results for the cognitive characteristics power-spectra (to evaluate the presence of autocorrelations) and powerlaw (to evaluate the presence of levy flight distributions). Shaded areas indicates values that lie within expected boundaries (Levy: agglomerated μ scores that lie within the range of [0,3] ; Autocorrelations using Fourier's power spectral transformations: α scores with expected boundaries between [-3, -.5]). * represent individual participant/model values that scored within the expected boundaries.

Sampling vs. Human results: Levy Flights

The powerlaw exponent was calculated by subtracting the difference between two consecutive numbers (b-a). In the direct sampler, the powerlaw exponents ranged between [2.42, 62.55]. The overall aggregated the time series had a μ score of 4.2. In total 2/50 samples in the DS produced a score that fits within a light-tailed distribution (R = -3.13, -5.6, $P < .001$). In the MCMC, the powerlaw exponents ranged between [14.81, 126.78]. The overall aggregated time series had a μ score of 76.48. In the MC^3, the powerlaw exponent ranged between [2.52, 119.86] with an aggregated score of 37.16. In total 7/50 samples produced a score that is a potential heavy-tailed distribution. 5/7 samples produced light-tailed distributions, $R \in [-2.407, -6.65], p < .01$ and the remaining two could not be determined ($R = -1.52, p = .12$ and $R = 1.34, p = .17$). The HMC produced powerlaw exponents between the range of [1.98, 85.9], with an aggregated score of 3.17. In total 5/50 samples produced a score that is a potential heavy-tailed distribution. Where most distributions were in favour of being exponential $R \in [-5.38, -2.72], p \in [.22, .83]$. This indicates that the MC^3, DS and HMC sampling model have the capacity to be similar to human powerlaw findings. Figure 4 details the μ values

To reiterate the experimental results, in the human condition condition, the powerlaw exponent ranged between [2.44, 68.4]. The overall aggregated time series had a μ score of 4.33. On average, all three models produced values outside a powerlaw distribution. It should be noted that the maximum range in all three models was twice as high as the human value. The MC^3 however did produce the occasional model that could be candidate for a heavy-tailed distribution and this is proportional to the human data. A 1 x 4 welch-anova determined little difference between the model and human data when evaluating powerlaw fits. The MC^3 model produced the largest overlap between the human level and a sampling one (t = -.912, p = .76), where the HMC had a p value of .52 and the MCMC had a p value of .49. This indicates the potential

fit of sampling in explaining the powerlaw exponents.

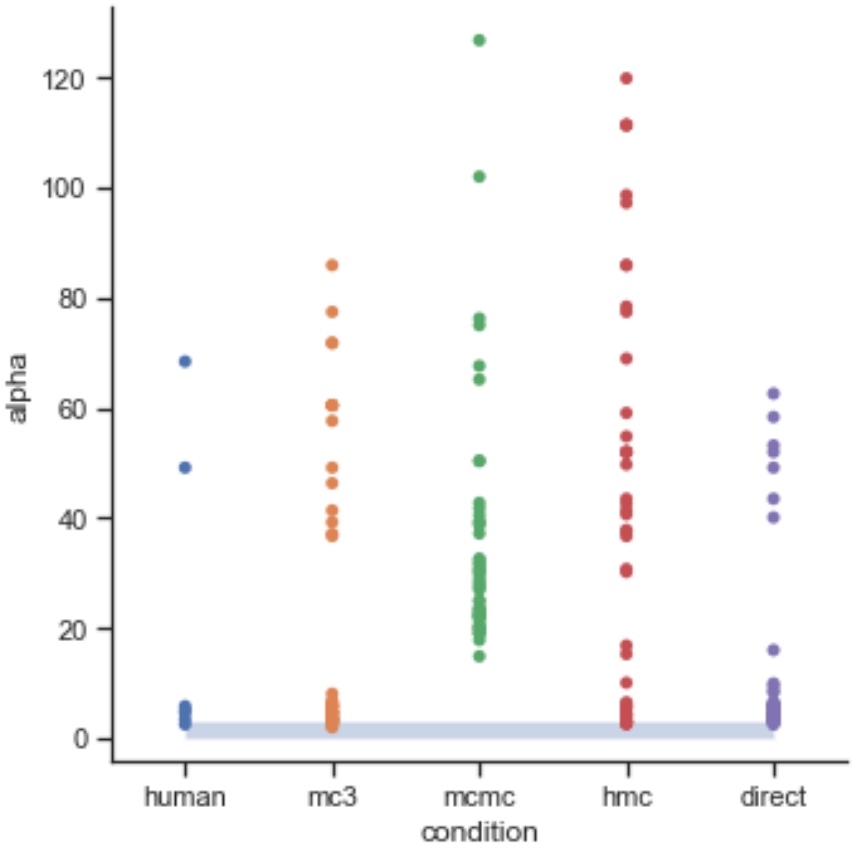


Figure 3: shows the μ results for the 4 samplers (Human, DS, MCMC, MC^3, HMC) and human data. The blue box represents values inside a powerlaw distribution indicating a potential heavy-tailed distribution

Sampling vs. Human results: Power Spectra

Figure 4 shows the mean power spectra value. In the direct sampler, the power-spectra exponents ranged between [-.56,.76], mean = .003. This indicates values between white noise and autocorrelatory behaviour. In the MCMC, the power-spectra exponents ranged between [-1.787, -.769], mean = -1.294. All values ranged between Brownian to pink noise. In the mc^3 model, the power-spectra exponent ranged between [-1.703,-.499], mean = -1.284. All values ranged from Brownian to near white-noise autocorrelations. The HMC model, produced exponents ranging from [-0.48,0.91], mean = 0.04. Similar to the $MC3$, the values also were spread from Brownian to white noise. A 1 x 4 welch-anova determined a difference between the models and the human data when assessing power-spectra exponents ($F(4,38.9) = 286.54, p < .001$). The direct sampler and HMC produced scores similar to the human results with medium effect sizes (DS: $t(7.16) = 1.874$, $p = .35$, $h = 0.746$; HMC: $t(7.649) = 1.54$, $p = 0.53$, $h = 0.61$). The remaining two models produced scores significantly different from the human data (MCMC: $t(7.571) = 13.46$; MC^3: $t(7.23) = 11.81$; ,$p < 0.001$. This indicates that given this task, humans are likely to use direct or HMC sampling.

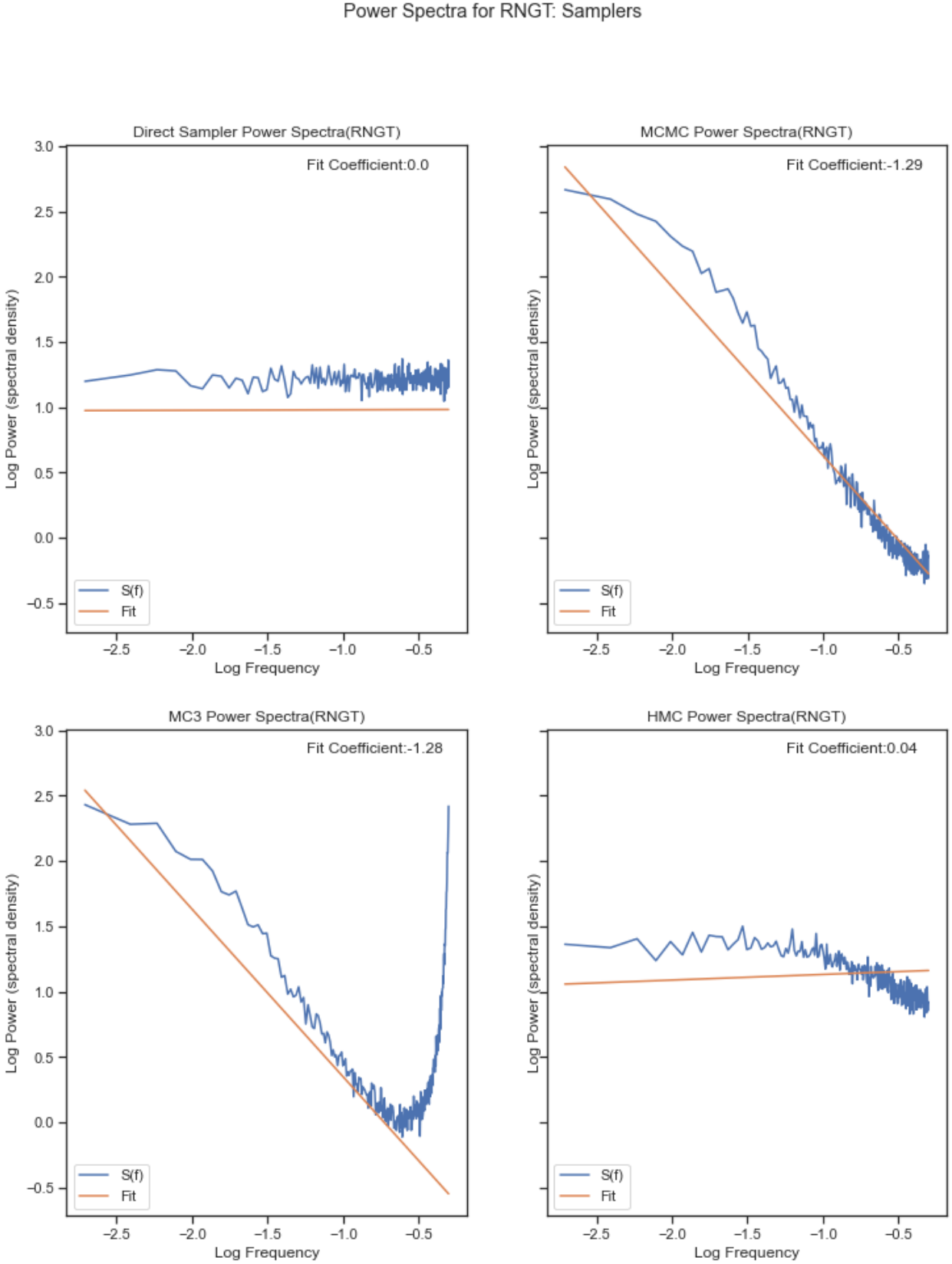


Figure 4: shows the power-spectra results for the 4 samplers: Top Left- DS; Top Right - MCMC; Bottom Left - MC^3; Bottom right - HMC. The graph is represented using a log-log plot. Each plot represents the mean power-spectra score taken from 50 models. The S(f) value represents the average frequency over time. The intercept value represents the mean coefficient output. Note: due to the variation in the number of responses per participant, the human average could not be represented

Sampling vs. Human results: Randomness

Evaluating randomness compares the models to the human data using Towse random criteria and the metrics. The results tell a different narrative to the previous cognitive characteristics. Means and standard deviations of all dependent measures are shown for the human data and the simulated models in table 5 shows the outcome.

Metric	**HUMAN**	**DS**	**MCMC**	**MC3**	**HMC**
A	.35 (.11)	.17(.01)	.44 (.027)	.20 (.040)	.0.13(.03)
RNG	.506(.04)	.676(.003)	.466(.012)	.676 (.04)	.54(.01)
R (%)	2.4 (2.2)	.39 (.173)	3.2 (1.6)	2.3 (1.1)	2.07(.89)
TPI	.81 (.15)	.95 (.035)	.48 (.031)	1.22 (.053)	.61 (.07)
Auto	.195 (.25)	-.007 (.05)	.918 (.013)	.03 (.18)	.23(.05)

Table 5: Means (and standard deviations) of the five dependent measures for the subject data of experimental data ("Human") and simulation results (Direct Sampling (DS), Markov chain Monte-Carlo (MCMC), Metropolis-coupled MCMC (MC^3), Hamiltonian Monte-Carlo (HMC)). Here A represents adjacency, RNG, equality of bigram use, R is the redundancy score, TPI represents the turning point index from the T&N paper and Auto represents the autocorrelation function from the Morariu et al (1995) paper.

A welchs anova determined that there were significant differences in all randomness variables (A; F(4,37.29) = 958.89, RNG; F(4,36.45) = 8925.02, R; (F(4,35.6) = 108.68, TP; F(4,37.5) = 2240.73, Auto; F(4,35.37)= 4960.79, $P < .0001$) between the various sampling algorithms $p < .0001$. All results can be found in the appendix 7.5. Post hoc analysis revealed preferences of certain algorithms over others. For the adjacency values (A), the MCMC (t(6.087) = 2.449, p =.28) produced a non-significant score when comparing the algorithms to the human data, to a small effect size, (.145). The RNG identified that the MC^3 and HMC was closest to the human data (MC^3: t(6.14) = 2.543, p = .12, HMC: t(7.720) = 2.20, p = .21) where all other values produced statistical significant results ($p < .05$). The effect size for the MC^3 and HMC large (1.012, .88 respectively). When evaluating the redundancy (R) value, MC^3, HMC and $MCMC$ produced the closest scores to the human dataset (MCMC:t(6.925) = .910, p = .86; HMC:t(6.27)= .389, p = .9 ;MC^3: t(6.397)= .132, p = .9). There was a micro effect size (.05) in the MC^3 and small effect size in the MCMC and HMC(.36, .155 respectively). The Turning point index (TP) was most similar to the Direct sampling algorithm with a large effect size of .97. All other results produced significant values ($p < .001$). Finally, the autocorrelation randomness score identified that the the DS, HMC and MC^3 produced similar values to the human dataset (DS: t(6.08)=2.14, p=.24; HMC: t(6.10)= .421, p = .9; MC^3: t(6.92)= 1.6 p=.5). The DS produced a large effect size (.85) and the MC^3 produced a medium effect size (.639) and the HMC produced a small effect size (.16). Only the MCMC produced significant differences from the human data $p < .001$

Frequency (%)	**HUMAN**	**DS**	**MCMC**	**MC3**	**HMC**
Min	75	97	46	86	85
Max	100	100	54	100	100
Mean	89.1	99.4	49.32	95.1	94.1

Table 6: mean, min and max of the total frequency (%) of the bigrams space per each sample algorithm. A 100% value indicates all bigram pairs have been produced and a 0% value indicates no bigram pairs have been produced

Table 6 outlines the difference of phase space used by each model. This shows the contrast between how the samplers explore their space. The direct sampling and MC^3 produces the largest exploration of space (average 99%) followed by the MC^3 (average 95.1%), followed by the HMC (94.1%) and then Human models (average 89.1%). The MCMC and HMC explored the least amount of space. Figure 6 shows the output of transitional matrix which also includes both dual and single task selected from the human participants. Subjectively, the transition matrix outlines the average strategy from step N_i to N_{i-1} for each model and participants (an overview for each participant can be found in the appendix 7.4). Visually, in both human conditions (Figure 6a and 6b), there appears to be a higher occurrence per sequential pairing of numbers (e.g. [1, 2] or [2,3]) and a lower occurrence of repeating the same number. There is also appears to be different

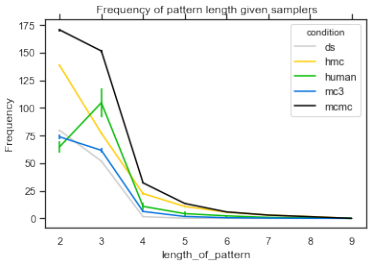
strategies where some participants are more stochastic and resembled a direct samplers output whilst others explored their space less effectively.

The patterns also draw out interesting comparisons between human and the model data. Overall, the patterns produced different 'favoured' algorithms when comparing to the human data for each feature observed.

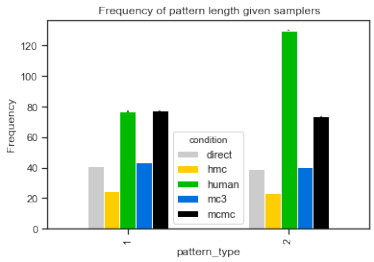
An ancova determined a difference between each model F(4,1650) = 70.03, $p < .001$. Except for a run length of two, the average run length per participant was much higher than a direct-sampling approach (see appendix 8 and figure 5a). The the run length produced overlap with the human data when using a DS, HMC and MC^3 (MC^3: t(57.626) = 1.01, p = .83.; DS: t(63.85) = 1.265, p = .69; HMC: t(78.88) = 1.535, p = .534). The MC^3, HMC, and DS both produced a small effect size (.144,.219,.18 respectively). The MCMC produced a different frequency output than the human single task (MCMC: t(106.34) = 3.81, $P < .05$).

When exploring the pattern types, an ancova also produced statistical result against the various samplers F(4,4979) = 95.14, $P < .001$. Only the MCMC model produced non-significant results to the human data (t(328.707) = .24). Albeit a small effect size (.135). The MC^3, HMC and DS produced statistically different scores from the human data (MC^3: t(308.42) = -5.307, HMC:t(325.94)= -2.893, DS: t(289.431) = -5.799, $P < .001$).

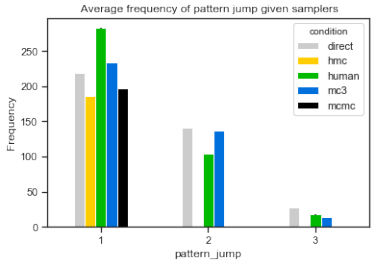
Lastly, when evaluating the jump lengths an ancova determined a difference between each model F(4,523) = 108.74 $p < .001$. Post hoc analysis revealed that the all four sampling methods were not statistically significant from the human data (MCMC: t(25.14) = 1.29, p = .67; MC^3: t (23.66) = .23, p = .9 ; DS: t(22.6)= .21, p = .9; HMC: t(20.82) = 2.63 , p = .08). HMC had the medium effect size (.61), the MCMC had a small effect size (.32) and the MC^3 and DS both had micro effect sizes of .05.



(a) Run length: [1,2,3] vs [1,2]

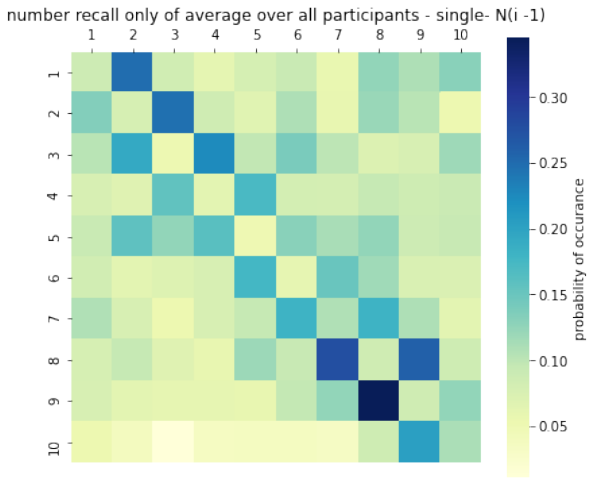


(b) Pattern type: [n+1,n+2,n+1] vs [n+1,n+2,n+3]

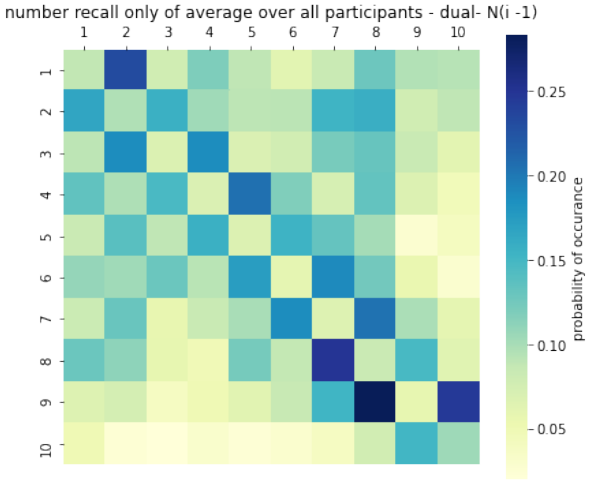


(c) Pattern jump: [1,2] vs [1,5] vs [1,8]

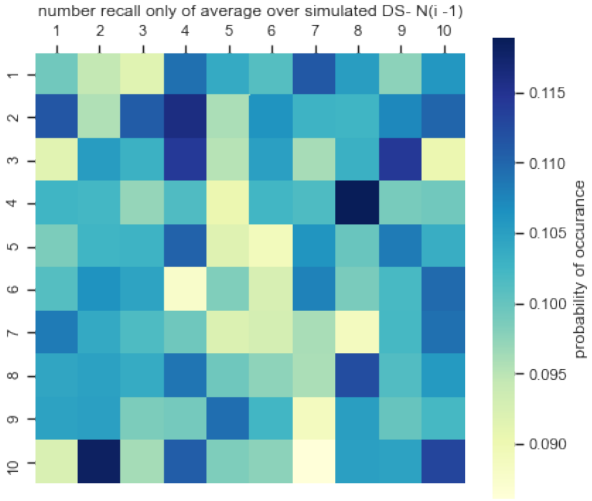
Figure 5: The three figures above show the various models performance. (a) demonstrates the run length of the pattern with an example of what the model compared to. (b) represents the pattern type selected and (c) represents the difference between pattern jump.



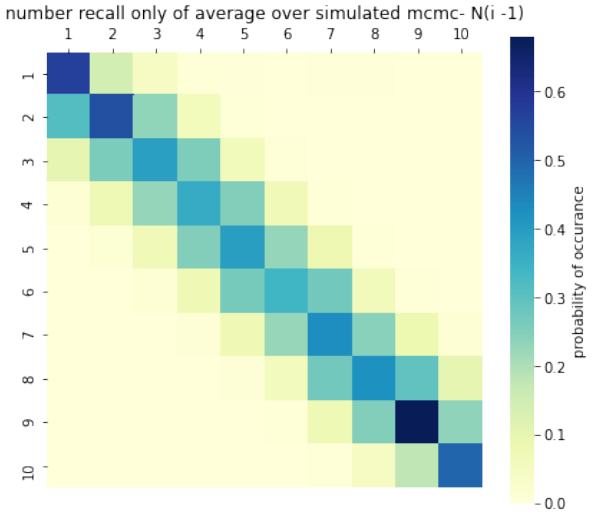
(a) (Human) Single Task



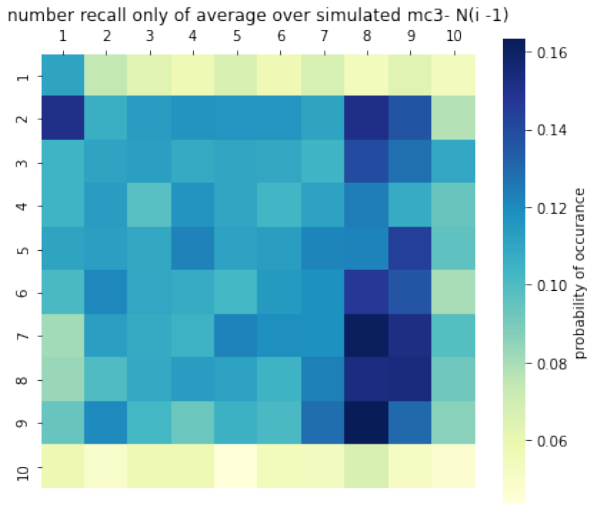
(b) (Human) Dual Task



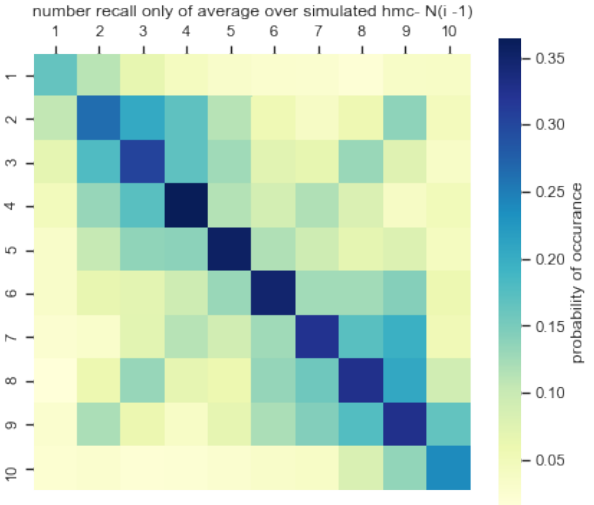
(c) (Model) Direct Sampling



(d) (Model) MCMC



(e) (Model) MC^3



(f) (Model) HMC

Figure 6: shows the average transitional matrix value over the 10 digits for N (y-axis) appearing given N-1 (x-axis). Darker blue colours represent higher probabilities while lower squares represent a lower probability. A true uniform distribution is expected to output a value of .11 for each digit.

6 Discussion

The goal of this thesis was to evaluate three main points. The first goal was to evaluate whether two foraging characteristics co-occurred when exploring RNGT and a time-interval estimation task. The answer to this is no. The findings indicate that ubiquitous cognitive characteristics, autocorrelations and levy distributions, do not co-occur but both effects can be found in the selected cognitive task. As shown in the results, values inside the range of powerlaw were more likely to appear in the random number generation task but not the tapping task and values which fit in the range of autocorrelations (1/f noise, Brownian motion) were more likely to be observed in the tapping task but not the random number generation task. The second point was to evaluate any distinguishing differences between a single and dual task. Dual tasking played more of a role in the tapping task than the random number generation. In fact, there were no difference between the single task and the dual task when exploring the randomness cognitive characteristic. Exploring the three cognitive characteristics using sampling algorithms beyond direct sampling seem promising but it is clear there is no defining algorithm capable of combining all three characteristics. The results highlights the importance of selecting tasks in identifying characteristics. The second important contribution is that different sampling algorithms represent different cognitive features given the human data. A non-significant statistical significance symbolised overlaps between human (single-task) and sampling model. In the random number generation task, direct sampling and the HMC produced results most similar to the human data in evaluating ubiquitous cognitive characteristics such as the powerlaw and autocorrelation and the sampling algorithms matched up to 3/5 metrics in the task-specific cognitive characteristics and 2/3 pattern outputs.

The next few sections discuss the results based on the characteristics and the further implications as well discusses some possible research expansions.

6.1 Ubiquitous Cognitive Characteristics

6.1.1 Levy-flights and Powerlaw

The tapping task did not produce outputs that resembled a heavy-tailed distributions and therefore did not produce levy flights. In the single task, this value was much higher than in the dual task. To the authors knowledge, evaluating a tapping task using powerlaw has not yet been investigated. One explanation to this contrast is that participants do not maintain an internal constant of time. One way to ensure a participant maintains an internal constant of time is to incorporate feedback into the response (Kuznetsov and Wallot, 2011). The single task did not require the participant to produce foraging behaviour thus it is expected that these values remain outside a foraging/ levy-like distribution. What is interesting is when incorporating a foraging related task in combination to the tapping task, the responses from this task moved much closer to a powerlaw distribution range albeit the results still remained outside the range of a heavy-tailed distribution. To some extent, this potentially indicates that a non-foraging task resembles foraging like behaviour when combined with a foraging task. Unfortunately as the RNGT did not always produce heavy-tailed distributions, it can not be conclusively determined but it would be interesting to follow up on whether this effect of the tapping task (or other non-foraging related tasks) also occurs when exploring to other memory generation tasks which have been proven to produce heavy distributions such as animal naming tasks or country recall (Patten et al., 2020).

Generally, the RNGT did not produce powerlaw exponents similar to heavy-tailed distributions. The exceptions of course is that 1/7 participants in the single task and 2/7 participants in the dual task produced powerlaw exponents. The goal of course was to generate a sequence of numbers as if pulling them out of a hat with replacement. One example caused by a lack of powerlaw exponents could be the fact that there is no known active strategy that can be incorporated unless the pattern itself become predictable. Consider the task of country recall. If the goal is to name all the countries, a participant might strategise by grouping a country using an alphabetical, continental, experiential or combination of the strategies provided. The jumps reflect the distance based on the grouping. On a number line, employing these strategies are less easy to categorise and/or represent. At most, the jumps between a country could resemble the jumps on a number line which was why incorporating more subjective pattern features (such as run length) was important to follow up on as a levy-flight distribution would favour a series of short jumps over long jumps. This indicates that for foraging behaviours, being able to evoke a strategy over a large, undefined distribution is more

important than evoking a strategy over a limited defined distribution. To test this assumption, increasing the set of variables (e.g having participants generate over 195 - the number of countries in the world), should produce more levy-like distributions or explore if a participant is able to identify a strategy (recalling evens vs odds or known sequences such as prime numbers). It is uncertain whether there is a large enough effect between the single and dual task due to the sample size of participants.

When exploring the RNGT task using the various sampling methods. Three particular algorithms stood out in producing similar output to the human (single-task) data. These were the DS algorithm, the HMC and the MC^3. These three algorithms are of particular interest because they were able to produce values that lay within the boundaries. The similarity between the DS and the MC^3 is that whatever the chain is exploring will usually be output as the acceptance is high. As a direct sampler is a random number generator, every value it generates will tend to follow a uniform distribution. This means that whether it generates a powerlaw distribution will also be decided randomly as each successive chain has the option to jump to a greater distance along the number-line. In a MC^3 algorithm, the exception to this are the edges of a set (1 and 10). The jumps between these chains are still situated in locality due to the number of chains currently in use. It would be expected that if the sample contained an infinite amount of parallel chains or a number of chains greater than the total set, it would resemble the same output as a DS algorithm. As the HMC applies metropolis principles to the MC^3, albeit with additional momentum, it explains how there are similarities in outputs. The number of chains were chosen arbitrarily due to the uniform distribution and limited set size. In this situation, having 2 chains resembled the same percentage as the human data (14%) for a possible heavy-tailed distribution. Further questions can evaluate the set size and chain lengths. If the set sizes increase, does the ability to simulate a heavy tailed distribution remain constant given the number? Does the output remain consistent when evaluating different sequence lengths or participant numbers?

6.1.2 Cognitive Characteristics: Power Spectra and Noise

In the tapping task, the behaviour produced autocorrelations similar to 1/f noise. This is similar to the 1/f noise findings found in the Zhu et al (2018) paper. The power spectra coefficients also echo similar results mentioned in their paper, where the coefficients ranged between [-0.9, -1.2], mean = -1.04 and in this research, the coefficients ranged between [0.3, 1.38], mean = -1.04. The effect has been described using the Wing and Kristofferson shifting strategy model where the execution of each tap is affected due to motor delay (Torre and Wagenmakers, 2009). An alternative explanation to the lag in the larger deviation could be due to the metronome intervals. The findings from Gilden et al., (1995) required participants to respond to a metronome 1 beat per second which may appear more natural and result in less deviations . An alternative explanation to the variation is that due to the motor delays, the output produces values closer to white noise. This is supported by the fact in both conditions, participants anticipated the presses a lot more quickly (on average 250ms quicker than hearing the beep) and their internal rhythm shifted to produce faster presses. The tapping task identified that autocorrelations (mean = -1.01) were reduced to white noise (-0.51) when introduced to a dual task environment. This indicates a shift in cognitive processing where an additional element became more demanding as explained in Alves et al., (2020) however, the overall performance of the RNGT did not differ. Participants now have to share their cognitive resources as proposed by a cognitive-sharing model. Sampling would explain this shift as the reduction of cognitive chains on the tapping task or that the increase in cognitive load produces a limited capacity in the chains. However to measure the effect of cognitive load would require modelling a sequential task where 1/f noise correlations occur simultaneously.

The results also confirm the results found in anti-correlation in RNGT tasks by Morariu et al (1999). In RNGT, Anti-correlation means that high numbers will often be followed by small numbers. Their paper produced a spectrum of $0.2 < \alpha < 0.5$, (mean = 0.24), whilst our findings found a larger range between $-0.19 < \alpha < 0.53$, (mean = 0.3) (V. V. Morariu, Coza, Chis, Isvoran, and Morariu, 2001). This indicates that sampling over a uniform distribution does not produce long-range nor short-ranged autocorrelations. Morariu et al. (2001) also addresses what happens at high set sizes where it indicates that 1/f scaling noise is produced by higher set sizes. A sampling approach explains this due to the limited set size and number of active chains. There are two chains sampling from the same/flat distributions, because the space is also much smaller, when a different chain makes a proposal (hot chain), the value will automatically be accepted and

results in more direct sampling than patterns. As highlighted the more chains involved in a sampler means the closer it is to resembling a direct sampler. The fact that this behaviour is anti-correlated could mean the semantic representation of a number line isn't how participants viewed the task. Future research could investigate the variations in parameters for how distance is measured. In this research, we viewed a number line linearly, but as we are conscious of the boundaries and set size, we instead treat the number line as a continuous value where the two edges connect and form an infinite loop. This would mean if a user jumped from 10 to 1, the difference would be 1 instead of 9 and vice versa. This indicates that returning to the beginning of a set is considered the underlying default representation as highlighted in Wulf, Hill & Hertwig (2020). An alternative hypothesis is that, when a set is smaller, creating wormholes to jump between distributions of memory could also explain the anti-correlated results. The difference between dual tasks and single tasks by comparing the cognitive characteristics makes it difficult to interpret the results. On one hand, the tapping task identifies differences from autocorrelations to white noise. This contributes to the underlying explanation in why humans become less efficient in parallel-like tasks - that is we have a limited number of chains which each have a limited capacity. However on the other hand, the random number generation task was still capable of producing anti-correlations where the dual task produces an pattern of high then low values values albeit not significantly. What is prominently contrasting is that the results explain both the impact caused from cognitive load (as observed with the heavy-tailed distribution in the tapping task) and not an impact in cognitive load (as observed with the autocorrelations generated from the random number generation task).

When exploring the power spectra and noise in the RNGT tasks using the sampling algorithms. The DS and the HMC produced a value that did not significantly differ from the human data set. What should be noted is that the remaining samplers produced a variety of ranges that averaged usually to 1/f noise which is considered an opposite effect of anti-correlated behaviour. For the MCMC, the chains are usually single which means the movement of travelling up and down is dependent on the location it is currently at. Naturally because of the locality, it is likely to see the same two orderings of a number repeat resulting in 1/f noise. Explaining this effect in the MC^3 is puzzling. It is expected that an MC^3 would not produce autocorrelated values as the probability of landing on the edge cases reduces the chance of jumping immediately from a 10 to a 1 and vice versa but for the remaining pairs, one would expect a more stochastic jump due to the high acceptance rate. One would expect a more random output as demonstrated in the HMC and because the HMC and MC^3 shares the same property in acceptance, it is surprising that this effect is different.

6.2 Task specific Cognitive Characteristics

In order for mental sampling to replicate autocorrelations resembling pink or Brownian noise, it appears that using large undefined sets where there are different multidimensional distributions (such as a Gaussian distribution) or use large spaces where the set size is large enough to cause chains to explore different environments to jump between. The randomness section is split into evaluating metrics already in place taken from previous papers as well as selected patterns to potentially identify differences between the algorithm.

6.2.1 Randomness Metrics

The use of metrics from various papers provides a future comparative score in the samplers as well as participants performance on the random number generation task for approximately 480-510 elements in a sequence. In general, most samplers produced scores, in the randomness metric, more towards the lower end of a spectrum [0 - .5] or 0% to %50. The exception to this was the autocorrelation function which produced values from 0 - .92. In general, the narrative of human randomness produces values that are between random and highly predictable and produced more biased values than random as shown when comparing the human results to a direct sampler. As most of our values explore the concept of producing stereotyped strings, and equipotentiality, what can be explained through the results is the use of variance produced in these values. In limited response sets, humans are able to explore their mental environment with relative ease, as shown in the R and autocorrelation metrics. Unsurprisingly, they are of course susceptible to biases (e.g. more likely to sequentially recall neighbouring numbers - as shown in the RNGT and A and are more likely to create momentum -as shown in the TP index) when evaluating how they move within that space which can be see when comparing the values to the direct sampler. We explore some of these biases, using three patterns, in

the next section. The scores when compared to other studies find similarities and differences. The Adjacency (0.328; Cooper's:0.345; our results) and TPI(0.73:0.81) values are very similar to the findings from Cooper but different in the RNGT(0.3: 0.5) and R (0.96: 2.4). The difference in values are most likely due to the task and length of sequence. As Cooper bases his findings from a sequence length of 100 letters which has been and the RNGT task itself [4]. The expected value of R decreases to zero as a sequence length increases to infinity due to the probability of independently recalling exactly the same number for each element in a set. Unlike Cooper's findings the bias values did not increase when placed under dual tasks which indicate a potential difference between bottleneck and parallel processing tasks. In this specific example, there was very little difference between the dual and singular task in random number generation. Interestingly, the autocorrelation function also identifies the lack of correlation similar to the power spectra values.

When comparing the metrics to the sampling algorithms and human data, the DS matched 2/5 criteria finding similar values in the Turning Point index and autocorrelation function. The MCMC produced identical values in 2/5 effects. It produced similar values in A and R. The MC^3 met 3/5 effects where it did not replicate similar values with the adjacency and turning point index. The HMC was able to meet 3/5 effects which is found in the RNG, autocorrelations,and R. The results indicate that a HMC and MC^3 model is a potentially good fit in simulating human number generation. The metric it did not perform well in is the autocorrelation function. This would indicate that whilst there is some relationship between neighbouring pairs and some momentum as identified with the TPI, the momentum is not predicted over continuous correlations in sequential behaviour (which is also confirmed through the power spectra values). In other words, generating values, one sample chain isn't enough. The results indicate that an MCMC sampler produces too much momentum due to its locality in generating chains but are capable of reproducing the biases to a human level. This indicates a need for a hybrid like model that is capable of reproducing the effects of random sampling (either a by combining a direct sampling with a HMC or a DS and an MC^3 or a HMC and an MC^3) whilst focusing on the biases that create these interesting effects without generating too much momentum as seen in the MCMC. Deviating into the various parameters and merging various algorithms models could explain these differences would be a natural extension in explaining mental sampling behaviour within restricted dimensionality.

6.2.2 Patterns

The aim of investigating various approaches in patterns was to gain further insight into the types of biases that are generated when undertaking random number generation and to investigate whether sampling was able to reproduce these effects. The single/dual task did not produce significant differences in any of the pattern analysis. The three patterns identified that the human participants were more likely to produce patterns of momentum in both the single and dual task conditions. For instance, participants, on average, produced run lengths of three over run lengths of two, and were also more likely to favour ascending/ descending pattern types over deviating ascending pairs. Lastly, participants were more likely to make short jumps over long jumps or medium jumps.

When examining how these algorithms move around the phase space, it is interesting to draw parallels between the HMC and MCMC. The transition matrix highlights the locality of the MCMC and HMC. In the MCMC, there is a bias to transition to neighbouring states. In the HMC, the momentum favours higher values. As the NUTs algorithm is the backbone of this algorithm, it adapts its stepsize setting based on the convergence value. Only a small burn-in was opted to assist in the convergence value of the step-size function however, there is still a large number of repeated values which could be due to a high rejection rate possibly due to the transformation of a uniform distribution created in the HMC algorithmic code. This suggests that in this use case, the HMC requires manual parameterisation in its implementation to be correctly evaluated. In the DS and MC^3 the movement of recollection is mostly stochastic. If we assume the simplest MC^3, where each chain can either go left or right on a number line, is currently located at 5, it is able to propose either a 4 or 6. At the edges, it can either stay at a 10 or go down to 9. The number of probabilities it can jump to jumps from 3 options to 2 options as it bounded by the set size. As the number of chains in

[4]Cooper's RNGT required participants to interact on a computerised clock interface and participants were asked to make the order of interface selection as random as possible

use was 2, it is possible for the algorithm to be situated in 2/10 places of an algorithm at a given location. When we evaluate both human trial, what is interesting to note is the combination of a DS model and an MCMC (or potentially a MC^3 model with more chains). The difference between the models and the humans is repetition avoidance. One might observe that combining a DS/ MC^3 model with a MCMC would potentially simulate similar result patterns but it is also important to identify an approach that includes avoidance.

When evaluating run length, the MC^3 and DS were able to create overlapping data. Descriptively, appendix 8 shows that human participants at most, produced run lengths of up to 8 and the MCMC, MC^3 and HMC also output values where this same effect occurred. The DS was able to produce 2/3 patterns, the MCMC produced What is interesting to note between the samplers and the human results is that all samplers were unable to replicate the distinguishing run differences between run length 2 and run length 3. This indicates different behaviour to both random number generation and a sampling methodology and raises questions about the driving mechanism behind this result. When exploring the difference between a sequential pattern and a deviation of a similar pattern, the MCMC was capable of reproducing similar results to the human data. As the focus of this paper investigated momentum, it is very likely that the MCMC will not always produce an accurate representation to human data when evaluating different patterns not explored in this paper. For instance, other pattern types would produce different results and so it is essential to explore different styles of pattern structures to understand further similarities and differences between samplers and human behaviour. All models produced non-significant results when comparing the output to the human data. The difference in numbers shown in Appendix 10 is due to the pattern excluding repeated values [3,3]. In certain models such as the DS, MC^3 and human outputs (as shown in the transition matrix), this difference isn't as large as the numbers are not so heavily dependent on repeated values however in models such as the HMC and MCMC, developing a model that is capable of avoiding repetition requires a different approach or a direct mechanism to avoid it. If samplers are able to reproduce repetition avoidance, expanding the analysis performed in this research could also extend to the repetition index.

6.3 Alternative explanations

Models to simulate random number generation has previously been explored using a Monte-Carlo combined with a Damerau-Levenshtein distance approach to explore patterns. In their findings, using a Markov rule of n-history within a Monte-Carlo system was believed to be the explanation of why their pattern matching system performed so well. In Appendix 7.4, we also show a confirmed visualisation of how different participants employ different heuristics when exploring their search space. Alternative approaches have used a mathematical model of random sequence generation using a Long short-term memory models with two main features Redundancy and a correlation function (Barbasz, Stettner, Wierzchoń, Piotrowski, and Barbasz, 2008). The papers aim was to simplify the analysis of randomness to two main features. However, explicit implementation of this approach has not been observed beyond its mathematical proof. Attempts to also explain the pseudo random number generation using chaos theory have also outlined the difficulties of explaining the biases elicited from human effects in generating such a sequence. Morariu et al. (1999), defined the result of a random number generation a combination of deterministic chaos where consciousness works with an imperfect algorithm. In Morariu's model, capturing self-avoidant behaviour was still difficult when exploring Shanons entropy. It seems even with an imperfect algorithm, developing such characteristics within an imperfect model remains much of a mathematical mystery beyond hard-coding specific heuristics that is capable of incorporating repetition avoidant behaviour.

6.4 Limitations

There have been a number of recommendations provided throughout the discussion that would help further understand the results in this paper yet there remains comments to be made on the experimental and algorithmic setup. For the experimental set-up, participants performed the experiment on their own machines which could have produced different deviations due to the computers performance. An online format was not chosen due to the variation in lag time of responses (Bridges, Pitiot, MacAskill, and Peirce, 2020). Thus, improving the experiments reliability would be beneficial for increasing exclusion criteria and internal validity reasons. Secondly, there are a couple of pattern results (such as the pattern jumps) that I remain sceptical

about and with more participants this can positively impact the statistical power and reliability of the results.

The algorithmic setup also contains its own limitations. For instance, there are a variety of implementations of chain swapping in the $MC3$. Here, we performed it for every iteration which is not noticeable when exploring two chains. However using this method for larger chain lengths will skew the way the MC3 works. For instance, if we had a MC^3 with a chain length of 10, the current algorithm would only make 1 chain swap per iteration where the expected algorithm ends up with 4-5 chain swaps by either: swapping two random pairs of chains together or randomly pair 2 neighbouring chains together and make a swap for each pair (as detailed in Zhu et al., 2018, Zhu et al., 2018a). Exploring random number generation in this setting using the different chain lengths would greatly validate the framework and appropriateness. The HMC NUTS model excludes the option to incorporate parameterisation. This indicates the need to investigate similar parameterisation options to reduce the high number of rejections. The exploration of HMC in cognitive decision making is relatively new so determining the right parameters and leapfrog implementation is also something of further interest.

6.5 Relevance to Artificial Intelligence

Cognitive models play an important role in bridging the fields within cognitive science and artificial intelligence. It bridges these two subjects by combining computational approaches in understanding the human mind. This project aims to use computational modelling, using Monte-Carlo algorithms to combine concepts at the computational and algorithmic level, as per Marr's level of analysis. To expand on Marr's level of analysis, the model is built within a rationally-bound cognitive framework meaning it aims to explore the human biases (A. N. Sanborn et al., 2010).

6.6 Conclusion

This thesis explored further diagnostic tools to evaluate three points : whether there is co-occurrence in foraging behaviours (autocorrelations and search paths) using an RNGT and tapping task, whether there is a difference in single and dual tasks and to explore (and compare) characteristic features of the random number generation task in both human participants and sampling algorithms to identify distinct behaviours. In this particular instance, the RNGT and a tapping task did not produce a co-occurrence of foraging behaviours. The only significant difference in the dual and single task was found in the tapping task. All other results found similar comparisons between the two tasks. The findings identified that the direct sampler, HMC or MC^3 is best when describing the foraging behaviours and an HMC/ MC^3 when exploring randomness features. The contribution identifies that there are mismatches between individual samplers and the features evaluated where further exploration of sampling models can combine task-specific characteristics and ubiquitous characteristics through a hybrid model or further parameterisation.

References

Alstott, J., Bullmore, E., & Plenz, D. (2014). Powerlaw: A python package for analysis of heavy-tailed distributions. *PloS one*, *9*(1), e85777.

Alves, M. V., Tassini, S., Aedo-Jury, F., & Bueno, O. (2020). Cognitive processing dissociation and mental effort manipulation in long demanding tasks. *BioRxiv*.

Anderson, J. R., Taatgen, N. A., & Byrne, M. D. (2005). Learning to achieve perfect timesharing: Architectural implications of hazeltine, teague, and ivry (2002).

Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language*, *85*, 27–41.

Arminger, G., & Muthén, B. O. (1998). A bayesian approach to nonlinear latent variable models using the gibbs sampler and the metropolis-hastings algorithm. *Psychometrika*, *63*(3), 271–300.

Baddeley, A. D. (1966). The capacity for generating information by randomization. *The Quarterly journal of experimental psychology*, *18*(2), 119–129.

Barbasz, J., Stettner, Z., Wierzchoń, M., Piotrowski, K., & Barbasz, A. (2008). How to estimate the randomness in random sequence generation task? *Polish Psychological Bulletin*, *39*(1).

Barthelemy, P., Bertolotti, J., & Wiersma, D. S. (2008). A lévy flight for light. *Nature*, *453*(7194), 495–498.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Brabazon, A., McGarraghy, S. et al. (2018). *Foraging-inspired optimisation algorithms*. Springer.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online.

Chevallier, A., Pion, S., & Cazals, F. (2018). Hamiltonian monte carlo with boundary reflections, and application to polytope volume calculations.

Cooper, R. P. (2016). Executive functions and the generation of "random" sequential responses: A computational account. *Journal of Mathematical Psychology*, *73*, 153–168.

Daniels, C., Witt, K., Wolff, S., Jansen, O., & Deuschl, G. (2003). Rate dependency of the human cortical network subserving executive functions during generation of random number series–a functional magnetic resonance imaging study. *Neuroscience Letters*, *345*(1), 25–28.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive psychology*, *96*, 1–25.

Figurska, M., Stańczyk, M., & Kulesza, K. (2008). Humans cannot consciously generate random numbers sequences: Polemic study. *Medical hypotheses*, *70*(1), 182–185.

Freidin, E., Aw, J., & Kacelnik, A. (2009). Sequential and simultaneous choices: Testing the diet selection and sequential choice models. *Behavioural processes*, *80*(3), 218–223.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G. (2006). Bounded and rational. In *Contemporary debates in cognitive science* (pp. 115–133). Blackwell.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart* (pp. 75–95). Oxford University Press.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, *8*(4), 296–301.

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological review*, *108*(1), 33.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f noise in human cognition. *Science*, *267*(5205), 1837–1839.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

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

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res. 15*(1), 1593–1623.

Humphries, N. E., Weimerskirch, H., & Sims, D. W. (2013). A new approach for objective identification of turns and steps in organism movement data relevant to random walk modelling. *Methods in Ecology and Evolution*, *4*(10), 930–938.

Jahanshahi, M., Profice, P., Brown, R. G., Ridding, M. C., Dirnberger, G., & Rothwell, J. C. (1998). The effects of transcranial magnetic stimulation over the dorsolateral prefrontal cortex on suppression of habitual counting during random number generation. *Brain: a journal of neurology*, *121*(8), 1533–1544.

Janczyk, M., & Kunde, W. (2020). Dual tasking from a goal perspective. *Psychological Review*.

Johnson, M. R., Higgins, J. A., Norman, K. A., Sederberg, P. B., Smith, T. A., & Johnson, M. K. (2013). Foraging for thought: An inhibition-of-return-like effect resulting from directing attention within working memory. *Psychological science*, *24*(7), 1104–1112.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 99–127). World Scientific.

Kee, Y. H., Chaturvedi, I., Wang, C. K. J., & Chen, L. H. (2013). The power of now: Brief mindfulness induction led to increased randomness of clicking sequence. *Motor Control*, *17*(3), 238–255.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in cognitive sciences*, *14*(5), 223–232.

Kuznetsov, N., & Wallot, S. (2011). Effects of accuracy feedback on fractal characteristics of time estimation. *Frontiers in integrative neuroscience*, *5*, 62.

Lai, Y., & Spanier, J. (2000). Adaptive importance sampling algorithms for transport problems. In *Monte-carlo and quasi-monte carlo methods 1998* (pp. 273–283). Springer.

Leon, P. (2020). Mc3-pnas-master. https://github.com/charlespwd/project-titlehttps://github.com/PabloLeon/mc3-pnas. GitHub.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems* (pp. 2690–2798).

Loetscher, T., Schwarz, U., Schubiger, M., & Brugger, P. (2008). Head turns bias the brain's internal random generator. *Current Biology*, *18*(2), R60–R62.

Maehara, Y., Saito, S., & Towse, J. N. (2019). Joint cognition and the role of human agency in random number choices. *Psychological research*, *83*(3), 574–589.

Montez, P., Thompson, G., & Kello, C. T. (2015). The role of semantic clustering in optimal memory foraging. *Cognitive science*, *39*(8), 1925–1939.

Morariu, V., & Morariu, S. (1995). Human brain as a random walk generator. part i. the exploration capabilities in the phase space. *Romanian Journal of Biophysics*, *5*, 1–8.

Morariu, V. V., Coza, A., Chis, M. A., Isvoran, A., & Morariu, L.-C. (2001). Scaling in cognition. *Fractals*, *9*(04), 379–391.

N. Towse, J., & Valentine, J. D. (1997). Random generation of numbers: A search for underlying processes. *European Journal of Cognitive Psychology*, *9*(4), 381–400.

Oomens, W., Maes, J. H., Hasselman, F., & Egger, J. I. (2015). A time series approach to random number generation: Using recurrence quantification analysis to capture executive behavior. *Frontiers in human neuroscience*, *9*, 319.

Orscheschek, F., Strobach, T., Schubert, T., & Rickard, T. (2019). Two retrievals from a single cue: A bottleneck persists across episodic and semantic memory. *Quarterly Journal of Experimental Psychology*, *72*(5), 1005–1028.

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as bayesian inference under extreme priors. *Cognitive psychology*, *102*, 127–144.

Patten, K. J., Greer, K., Likens, A. D., Amazeen, E. L., & Amazeen, P. G. (2020). The trajectory of thought: Heavy-tailed distributions in memory foraging promote efficiency. *Memory & Cognition*, 1–16.

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of neuroscience methods*, *162*(1-2), 8–13.

Peng, C.-K., Mietus, J., Hausdorff, J., Havlin, S., Stanley, H. E., & Goldberger, A. L. (1993). Long-range anticorrelations and non-gaussian behavior of the heartbeat. *Physical review letters*, *70*(9), 1343.

Rakitin, B. C., Gibbon, J., Penney, T. B., Malapani, C., Hinton, S. C., & Meck, W. H. (1998). Scalar expectancy theory and peak-interval timing in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(1), 15.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., & Chong, S. (2011). On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking*, *19*(3), 630–643.

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as lévy foraging. *Physica A: Statistical Mechanics and its Applications*, *385*(1), 255–260.

Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and cognition*, *112*, 98–101.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, *20*(12), 883–893.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Sanborn, A., & Griffiths, T. L. (2008). Markov chain monte carlo with people. In *Advances in neural information processing systems* (pp. 1265–1272).

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: A systematic review. *BMC medical informatics and decision making*, *16*(1), 138.

Schulz, M.-A. [Marc-Andre], Baier, S., Böhme, B., Bzdok, D., & Witt, K. (2020). A cognitive fingerprint in human random number generation.

Schulz, M.-A. [Marc-André], Schmalbach, B., Brugger, P., & Witt, K. (2012). Analysing humanly generated random number sequences: A pattern-based approach. *PloS one*, *7*(7), e41531.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic bulletin & review*, *17*(4), 443–464.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Torre, K., & Wagenmakers, E.-J. (2009). Theories and models for 1/fβ noise in human movement science. *Human movement science*, *28*(3), 297–318.

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/f α noise in human cognition. *Psychonomic bulletin & review*, *11*(4), 579–615.

Ward, L. M., & Greenwood, P. E. (2007). 1/f noise. *Scholarpedia*, *2*(12), 1537.

Wulff, D. U., Hills, T., & Hertwig, R. (2020). Memory is one representation not many: Evidence against wormholes in memory.

Yuan, C., & Druzdzel, M. J. (2006). Importance sampling algorithms for bayesian networks: Principles and performance. *Mathematical and Computer Modelling*, *43*(9-10), 1189–1207.

Zhu, J., Sanborn, A., & Chater, N. (2018a). Mental sampling in multimodal representations. In *Advances in neural information processing systems* (pp. 5748–5759).

Zhu, J., Sanborn, A., & Chater, N. (2018b). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments.

7 Appendix

7.1 Justification for the types of tasks selected

Number generation tasks have been used to explore mental patterns in autocorrelations (**Morariu1999part3**), randomness (N. Towse and Valentine, 1997), dual tasking (Cooper, 2016) but in human participants, the distribution has not been investigated to evaluate the optimal foraging behaviours. Random number generation requires strategy in the selection and inhibition of responses. Participants are required to remember a base set and incorporate it into their own interpretation of randomness (Jahanshahi et al., 1998). As a generation task has been selected, in random generation tasks, evaluating the randomness of behaviour is difficult to define unless the scope has been specified. Using a specified scope allows sufficient constraints to compare identifiable characteristics thus number generation seemed like a good starting place. A more natural example would be free recall such as countries or animal names. In free recall, the ultimate goal is to generate as many items within a semantic category over a time. However, approaching the problem using a semantic category severely impacts a task due to its bottleneck effect as it requires activation and continuous moderation (Anderson, Taatgen, and Byrne, 2005; Janczyk and Kunde, 2020). Anderson et al., (2015) noted that only when the task was highly practiced and automatised would the bottleneck effect dissipate and this has recently been replicated (Orscheschek, Strobach, Schubert, and Rickard, 2019). Time estimation tasks have shown the ability to perform a secondary task in parallel (Rakitin et al., 1998). Time estimation tasks also incorporate time-series providing consistency needed to elicit autocorrelatory signal processing responses and have illicited levy-distribution (Gilden, 2001). For this purpose we combined two experiments from Zhu et al., (2018) that enable a capacity sharing model. For this reason, we can in effect combine the random number generation (RNG) task and interval tapping task into a joint goal rather than two distinct goal to potentially reduce and / or eliminates the task cost. By having participants represent a rhythm using interval estimation, the participants will then be able to layer multiple tasks without hindering the data collection too significantly. Comparing randomness scores has also proven useful when exploring the difference between singular tasks to dual tasks (Cooper, 2016).

7.2 Descriptive results

Participant	Single task			Dual Task	
	RNGT Sequence Length	IRI Sequence Length	RT (s)	RNGT + IRI Sequence Length	RT (s)
0	521	549	0.503	559	0.533
1	478	620	0.580	517	0.675
2	509	549	0.610	572	0.54
3	509	488	0.652	310	1.14
4	514	682	0.462	536	0.53
5	441	569	0.462	340	0.97
6	445	681	0.380	482	0.53
Mean	**488.14**	**591.14**	**0.52**	**473**	**0.7**

Table 7: Table shows the descriptive analysis (sequence length and mean reaction time) for all participants given the single and dual task. Mean scores for all participants is given in the last row of the table.

7.3 Pattern results

Length of Pattern	Mean Frequency Score				
	Human	**DS**	**MCMC**	**MC3**	**HMC**
2	64.86 (13.63)	79.68 (7.1)	170.72 (8.88)	73.84 (13.71)	153.78 (6.78)
3	104.71 (33.86)	51.86 (7.1)	151.60 (9.09)	61.62 (12.70)	131.86 (9.35)
4	10.86(8.82)	1.56 (1.05)	32.16 (4.28)	6.26 (2.66)	25.74 (3.46)
5	4.29(4.86)	0.14 (0.35)	13.48 (3.59)	1.72 (1.24)	9.92 (2.43)
6	2.29(2.56)	0.02 (0.14)	5.86 (2.35)	0.58 (0.73)	4.48(1.84)
7	0.86(1.46)	0 (0)	2.98 (1.71)	0.26 (0.49)	1.92 (1.44)
8	0.57(0.79)	0 (0)	1.54 (1.22)	0.1 (0.36)	0.98 (1.08)
9	0	0 (0)	0 (0)	0 (0)	0 (0)

Table 8: shows the mean (and standard deviation) frequency score across the various samplers for the run length. Human in this case represents the single task only.

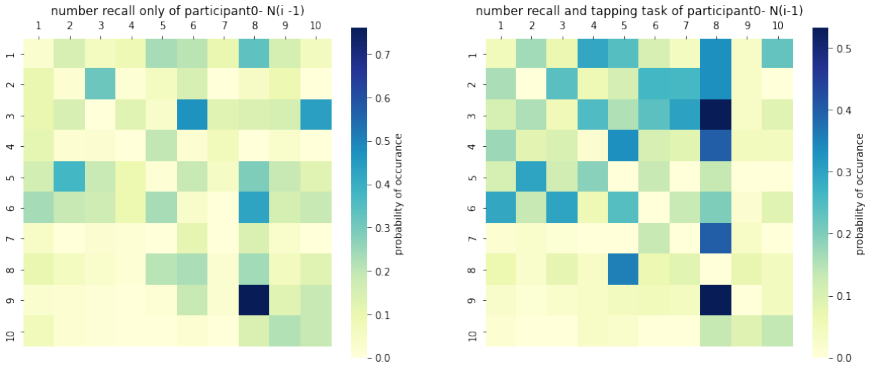
Pattern Type	Condition	Mean Frequency
1 (N+1, N+2.., N+i-1)	**HUMAN**	76.86 (22.51)
	DIRECT	41.02 (5.42)
	MCMC	77.20 (9.26)
	MC3	43.52 (9.26)
	HMC	24.54 (7.46)
2 (N+1,N+2,..,Ni)	**HUMAN**	130.00 (70.87)
	DIRECT	38.98 (5.50)
	MCMC	73.48 (8.86)
	MC3	40.38 (9.42)
	HMC	23.28 (7.69)

Table 9: shows the mean (and standard deviation) after first summing over the individual participants and then taking the mean of the condition and pattern types

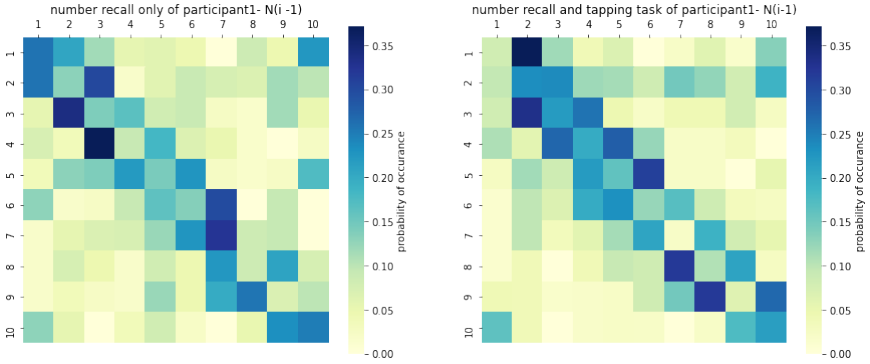
Condition	Jump_length	Frequency
Human	Short Range	283
	Medium Range	104
	Long Range	17
Direct	Short Range	219
	Medium Range	141
	Long Range	28
MCMC	Short Range	197
	Medium Range	1
	Long Range	0
MC3	Short Range	235
	Medium Range	137
	Long Range	14
HMC	Short Range	115
	Medium Range	68
	Long Range	15

Table 10: shows the mean (and standard deviation) after first summing over the individual participants and then taking the mean of the condition and jump ranges. All repeated values in the sequence have been removed thus indicates an uneven sequence length less than 512.

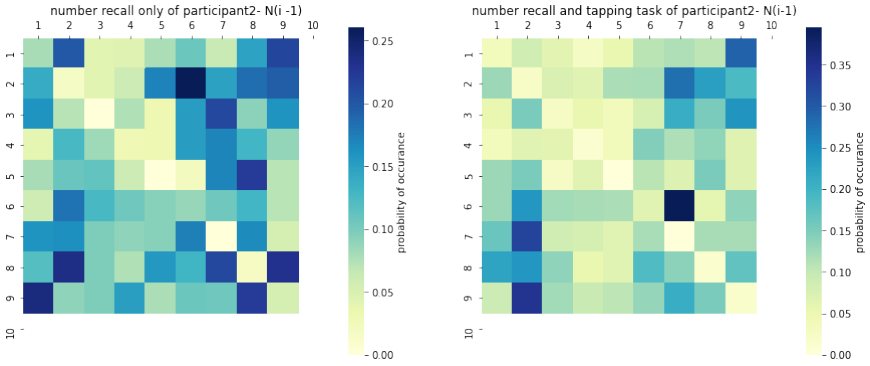
7.4 Individual Transition Matrices



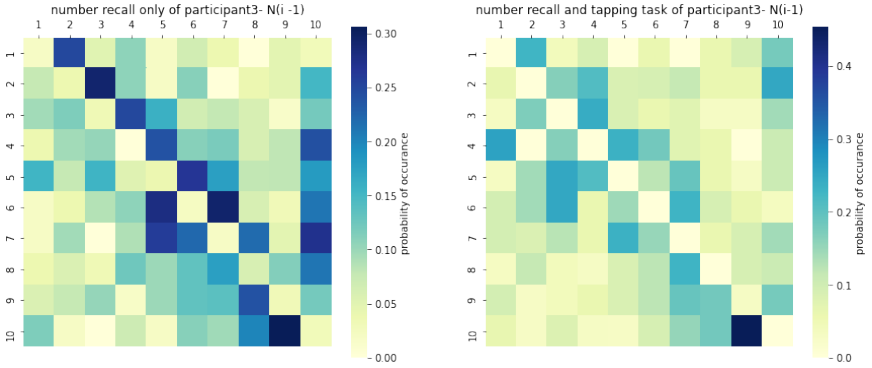
(a) Participant 0



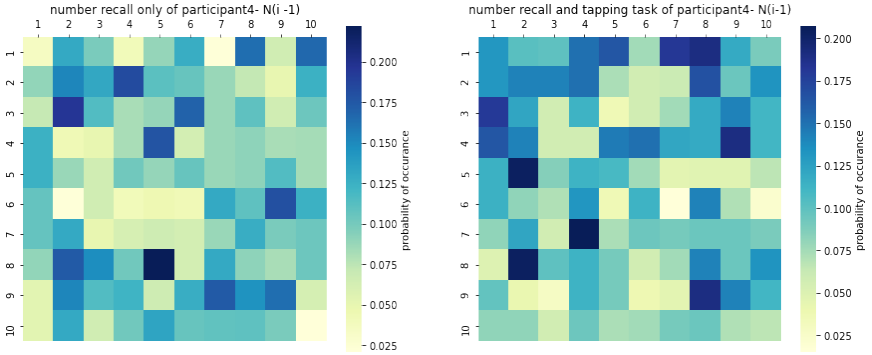
(b) Participant 1



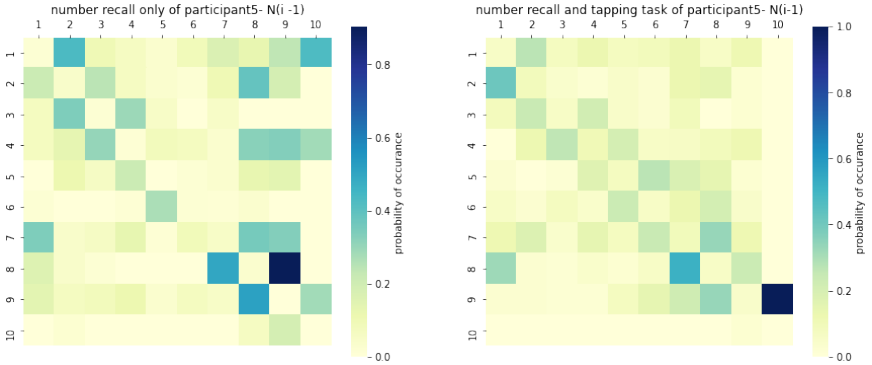
(c) Participant 2



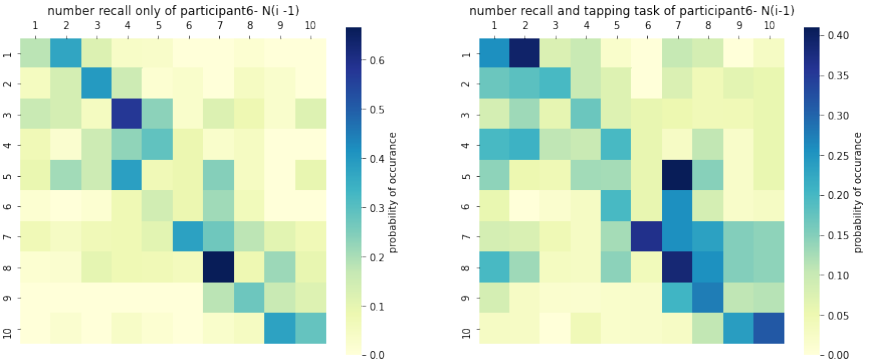
(d) Participant 3



(e) Participant 4



(f) Participant 5



(g) Participant 6

Figure 7: Figures show individual participants transition matrices. On the left shows the RNGT for the single task, and on the right shows when a tapping component is added. As shown the patterns generated are different from other participants however characteristics tend to remain.

7.5 Statistical tables

Note this section excludes t-test.

7.5.1 Foraging characteristics results

Levene's statstical analysis for powerlaw T-test:
$LeveneResult(statistic = 1.718931392472485, pvalue = 0.2143651413221265)$

Source	ddof1	ddof2	F	p-unc
condition	4	38.637	11.111	0.000004

Table 11: Table shows welchs

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
direct	hmc	11.560	37.637	-26.077	4.010	two-sided	-4.598	68.315	0.001000	-0.913
direct	**human**	**11.560**	**19.812**	**-8.251**	**7.456**	**two-sided**	**-0.783**	**6.636**	**0.900000**	**-0.311**
direct	mc3	11.560	20.087	-8.527	3.066	two-sided	-1.967	83.488	0.285240	-0.390
direct	mcmc	11.560	34.938	-23.378	2.747	two-sided	-6.019	91.184	0.001000	-1.194
hmc	**human**	**37.637**	**19.812**	**17.826**	**8.136**	**two-sided**	**1.549**	**9.339**	**0.523339**	**0.617**
hmc	mc3	37.637	20.087	17.550	4.472	two-sided	2.775	88.160	0.046384	0.551
hmc	mcmc	37.637	34.938	2.699	4.260	two-sided	0.448	80.231	0.900000	0.089
human	**mc3**	**19.812**	**20.087**	**-0.276**	**7.714**	**two-sided**	**-0.025**	**7.593**	**0.900000**	**-0.010**
human	**mcmc**	**19.812**	**34.938**	**-15.127**	**7.593**	**two-sided**	**-1.409**	**7.133**	**0.596876**	**-0.561**
mc3	mcmc	20.087	34.938	-14.851	3.386	two-sided	-3.102	95.494	0.017811	-0.616

Table 12: Table shows post hoc analysis for the powerlaw results using the analysis of alpha (μ) values. A B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler.

Source	ddof1	ddof2	F	p-unc
condition	4	38.897	286.539	2.747331e-28

Table 13: Table shows Welchs anova for powerlaw in a time series. Condition represents the four sampler models and human data

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
0	direct	hmc	0.003	0.043	-0.040	0.041	two-sided	-0.682	95.403	0.900000
-0.135										
direct	**human**	**0.003**	**0.245**	**-0.242**	**0.091**	**two-sided**	**-1.874**	**7.165**	**0.351064**	**-0.746**
direct	mc3	0.003	-1.284	1.287	0.038	two-sided	23.764	97.915	0.001000	4.716
direct	mcmc	0.003	-1.294	1.298	0.036	two-sided	25.367	97.243	0.001000	5.034
hmc	**human**	**0.043**	**0.245**	**-0.202**	**0.093**	**two-sided**	**-1.540**	**7.649**	**0.527681**	**-0.613**
hmc	mc3	0.043	-1.284	1.327	0.042	two-sided	22.441	96.215	0.001000	4.454
hmc	mcmc	0.043	-1.294	1.338	0.040	two-sided	23.719	92.253	0.001000	4.707
human	**mc3**	**0.245**	**-1.284**	**1.529**	**0.092**	**two-sided**	**11.811**	**7.238**	**0.001000**	**4.701**
human	**mcmc**	**0.245**	**-1.294**	**1.540**	**0.091**	**two-sided**	**12.005**	**6.970**	**0.001000**	**4.778**
mc3	mcmc	-1.284	-1.294	0.010	0.037	two-sided	0.201	96.668	0.900000	0.040

Table 14: Table shows post-hoc results for the power spectra, to determine the noise in time-series data, using a pairwise gameshowll analysis. A B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

7.5.2 Randomness

Model	Metric	Source	ddof1	ddof1	F	P-unc
Human	**A**	Dual vs. Single task	1	11.872	0.09	0.769423
	RNG	Dual vs. Single task	1	12.0	0.772	0.396879
	R	Dual vs. Single task	1	11.919	0.033	0.859178
	TP	Dual vs. Single task	1	11.999	0.196	0.66549
	AUTO	Dual vs. Single task	1	11.984	0.096	0.7626
Sampler	**A**	MCMC vs. MC3 vs. HMC vs. DS vs. Human	4	37.294	958.886	4.895642e-37
	RNG	MCMC vs. MC3 vs. HMC vs. DS vs. Human	4	36.459	8925.015	5.673911e-54
	R	MCMC vs. MC3 vs. HMC vs. DS vs. Human	4	35.572	108.683	1.988835e-19
	TP	MCMC vs. MC3 vs. HMC vs. DS vs. Human	4	37.489	2240.727	4.737926e-44
	AUTO	MCMC vs. MC3 vs. HMC vs. DS vs. Human	4	37.489	2240.727	4.737926e-44

Table 15: Table shows the results from 10 individual welchs-anova test. The table is split into two individual evaluation where the conditions for human represents the differences between the dual and single task, and the condition for sampler represents the differences between the models selected.

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
ds	hmc	0.179	0.132	0.047	0.003	two-sided	9.720	76.640	0.001000	1.929
ds	**human**	**0.179**	**0.345**	**-0.167**	**0.031**	**two-sided**	**-3.752**	**6.033**	**0.009029**	**-1.493**
ds	mc3	0.179	0.209	-0.030	0.004	two-sided	-4.918	65.211	0.001000	-0.976
ds	mcmc	0.179	0.436	-0.258	0.003	two-sided	-58.117	81.666	0.001000	-11.534
hmc	**human**	**0.132**	**0.345**	**-0.213**	**0.032**	**two-sided**	**-4.787**	**6.108**	**0.001000**	**-1.905**
hmc	mc3	0.132	0.209	-0.077	0.005	two-sided	-10.900	90.418	0.001000	-2.163
hmc	mcmc	0.132	0.436	-0.304	0.004	two-sided	-53.955	96.918	0.001000	-10.708
human	**mc3**	**0.345**	**0.209**	**0.137**	**0.032**	**two-sided**	**3.055**	**6.196**	**0.044181**	**1.216**
human	**mcmc**	**0.345**	**0.436**	**-0.091**	**0.031**	**two-sided**	**-2.042**	**6.087**	**0.279112**	**-0.813**
mc3	mcmc	0.209	0.436	-0.227	0.005	two-sided	-33.508	85.436	0.001000	-6.650

Table 16: Table shows post-hoc results for the A metric using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
ds	hmc	0.440	0.543	-0.104	0.004	two-sided	-17.839	49.814	0.001000	-3.540
ds	**human**	**0.440**	**0.506**	**-0.067**	**0.011**	**two-sided**	**-4.245**	**6.014**	**0.002523**	**-1.690**
ds	mc3	0.440	0.466	-0.027	0.001	two-sided	-14.766	58.146	0.001000	-2.931
ds	mcmc	0.440	0.676	-0.236	0.001	two-sided	-193.519	70.372	0.001000	-38.407
hmc	**human**	**0.543**	**0.506**	**0.037**	**0.012**	**two-sided**	**2.205**	**7.720**	**0.210559**	**0.878**
hmc	mc3	0.543	0.466	0.077	0.004	two-sided	12.774	57.578	0.001000	2.535
hmc	mcmc	0.543	0.676	-0.133	0.004	two-sided	-22.498	52.541	0.001000	-4.465
human	**mc3**	**0.506**	**0.466**	**0.040**	**0.011**	**two-sided**	**2.543**	**6.144**	**0.122290**	**1.012**
human	**mcmc**	**0.506**	**0.676**	**-0.170**	**0.011**	**two-sided**	**-10.757**	**6.059**	**0.001000**	**-4.282**
mc3	mcmc	0.466	0.676	-0.210	0.001	two-sided	-102.719	83.405	0.001000	-20.386

Table 17: Table shows post-hoc results for the RNG metric using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
ds	hmc	0.391	2.072	-1.682	0.091	two-sided	-13.104	52.701	0.001000	-2.601
ds	**human**	**0.391**	**2.401**	**-2.010**	**0.590**	**two-sided**	**-2.407**	**6.010**	**0.156209**	**-0.958**
ds	mc3	0.391	2.289	-1.898	0.108	two-sided	-12.428	51.587	0.001000	-2.467
ds	mcmc	0.391	3.188	-2.797	0.162	two-sided	-12.194	50.131	0.001000	-2.420
hmc	**human**	**2.072**	**2.401**	**-0.328**	**0.597**	**two-sided**	**-0.389**	**6.276**	**0.900000**	**-0.155**
hmc	mc3	2.072	2.289	-0.216	0.139	two-sided	-1.100	95.007	0.779001	-0.218
hmc	mcmc	2.072	3.188	-1.115	0.184	two-sided	-4.281	76.355	0.001000	-0.850
human	**mc3**	**2.401**	**2.289**	**0.112**	**0.600**	**two-sided**	**0.132**	**6.397**	**0.900000**	**0.053**
human	**mcmc**	**2.401**	**3.188**	**-0.787**	**0.612**	**two-sided**	**-0.910**	**6.925**	**0.860131**	**-0.362**
mc3	mcmc	2.289	3.188	-0.899	0.193	two-sided	-3.289	84.948	0.009877	-0.653

Table 18: Table shows post-hoc results for the R metric using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
ds	hmc	0.951	0.619	0.332	0.008	two-sided	29.942	72.249	0.001000	5.942
ds	**human**	**0.951**	**0.814**	**0.136**	**0.040**	**two-sided**	**2.434**	**6.096**	**0.148464**	**0.969**
ds	mc3	0.951	1.224	-0.273	0.006	two-sided	-30.266	84.996	0.001000	-6.007
ds	mcmc	0.951	0.485	0.465	0.005	two-sided	70.342	96.343	0.001000	13.960
hmc	**human**	**0.619**	**0.814**	**-0.196**	**0.040**	**two-sided**	**-3.452**	**6.384**	**0.017679**	**-1.374**
hmc	mc3	0.619	1.224	-0.605	0.009	two-sided	-48.632	91.407	0.001000	-9.652
hmc	mcmc	0.619	0.485	0.133	0.008	two-sided	12.335	67.307	0.001000	2.448
human	**mc3**	**0.814**	**1.224**	**-0.409**	**0.040**	**two-sided**	**-7.271**	**6.220**	**0.001000**	**-2.894**
human	**mcmc**	**0.814**	**0.485**	**0.329**	**0.040**	**two-sided**	**5.881**	**6.074**	**0.001000**	**2.341**
mc3	mcmc	1.224	0.485	0.738	0.006	two-sided	84.925	78.599	0.001000	16.855

Table 19: Table shows post-hoc results for the TPI metric using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
ds	hmc	-0.007	0.235	-0.242	0.008	two-sided	-21.539	96.415	0.001000	-4.275
ds	**human**	**-0.007**	**0.195**	**-0.202**	**0.067**	**two-sided**	**-2.141**	**6.075**	**0.240325**	**-0.852**
ds	mc3	-0.007	0.039	-0.046	0.019	two-sided	-1.708	57.172	0.431653	-0.339
ds	mcmc	-0.007	0.918	-0.925	0.005	two-sided	-120.971	55.100	0.001000	-24.009
hmc	**human**	**0.235**	**0.195**	**0.040**	**0.067**	**two-sided**	**0.421**	**6.097**	**0.900000**	**0.168**
hmc	mc3	0.235	0.039	0.197	0.019	two-sided	7.290	59.526	0.001000	1.447
hmc	mcmc	0.235	0.918	-0.683	0.006	two-sided	-79.048	53.721	0.001000	-15.688
human	**mc3**	**0.195**	**0.039**	**0.157**	**0.069**	**two-sided**	**1.606**	**6.915**	**0.492245**	**0.639**
human	**mcmc**	**0.195**	**0.918**	**-0.723**	**0.067**	**two-sided**	**-7.676**	**6.005**	**0.001000**	**-3.055**
mc3	mcmc	0.039	0.918	-0.880	0.018	two-sided	-34.267	49.514	0.001000	-6.801

Table 20: Table shows post-hoc results for the Autocorrelation metric using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
ds	hmc	0.391	0.720	-0.330	0.039	two-sided	-5.977	331.530	0.001000	-0.534
ds	**human**	**0.391**	**0.852**	**-0.462**	**0.149**	**two-sided**	**-2.192**	**34.690**	**0.190772**	**-0.395**
ds	mc3	0.391	0.845	-0.455	0.045	two-sided	-7.083	308.267	0.001000	-0.633
ds	mcmc	0.391	1.141	-0.750	0.058	two-sided	-9.088	283.517	0.001000	-0.812
hmc	**human**	**0.720**	**0.852**	**-0.132**	**0.153**	**two-sided**	**-0.612**	**38.127**	**0.900000**	**-0.110**
hmc	mc3	0.720	0.845	-0.125	0.056	two-sided	-1.575	483.859	0.733501	-0.141
hmc	mcmc	0.720	1.141	-0.420	0.067	two-sided	-4.438	423.305	0.001000	-0.396
human	**mc3**	**0.852**	**0.845**	**0.007**	**0.154**	**two-sided**	**0.033**	**39.886**	**0.900000**	**0.006**
human	**mcmc**	**0.852**	**1.141**	**-0.288**	**0.159**	**two-sided**	**-1.285**	**44.445**	**0.673449**	**-0.231**
mc3	mcmc	0.845	1.141	-0.296	0.071	two-sided	-2.949	464.538	0.026771	-0.263

Table 21: Table shows post-hoc results for all 5 metrics (A, R, RNG, TP, Auto) metric using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

7.5.3 Pattern results

Condition	W	pval	normal
single_frequency	0.657608	3.668430e-10	False
dual_frequency	0.652994	3.058511e-10	False

Table 22: Table shows an example of the pattern run length not meeting the assumption for normality.

Source	SS	DF	F	p-unc
condition	135.080000	1	0.188894	6.646983e-01
pattern_length	78327.520833	1	109.531869	3.669399e-18
Residual	77947.175167	109		

Table 23: Table shows ancova results for the pattern run length. Condition represents the whether it is a dual or single task results. Pattern_length represents 9 cases of pattern run length up to 9.

Source	SS	DF	F	p-unc
condition	2.484375e+05	4	70.031501	7.905273e-55
length_of_pattern	1.986885e+06	1	2240.314601	1.257865e-309
Residual	1.463348e+06	1650		

Table 24: Table shows ancova results for the pattern run length. Condition represents the 4 samplers + human single results. Pattern_length represents 9 cases of pattern run length up to 9.

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
ds	hmc	16.658	32.377	-15.720	1.962	two-sided	-5.665	670.194	0.001	-0.400
ds	**human**	**16.658**	**23.554**	**-6.896**	**3.853**	**two-sided**	**-1.265**	**63.951**	**0.685**	**-0.180**
ds	mc3	16.658	18.048	-1.390	1.479	two-sided	-0.665	797.945	0.900	-0.047
ds	mcmc	16.658	47.292	-30.635	2.584	two-sided	-8.384	548.134	0.001	-0.592
hmc	**human**	**32.377**	**23.554**	**8.824**	**4.066**	**two-sided**	**1.535**	**78.880**	**0.534**	**0.219**
hmc	mc3	32.377	18.048	14.330	1.967	two-sided	5.152	673.510	0.001	0.364
hmc	mcmc	32.377	47.292	-14.915	2.891	two-sided	-3.648	716.101	0.002	-0.258
human	**mc3**	**23.554**	**18.048**	**5.506**	**3.856**	**two-sided**	**1.010**	**64.106**	**0.829**	**0.144**
human	**mcmc**	**23.554**	**47.292**	**-23.739**	**4.400**	**two-sided**	**-3.815**	**106.341**	**0.001**	**-0.543**
mc3	mcmc	18.048	47.292	-29.245	2.587	two-sided	-7.993	550.452	0.001	-0.565

Table 25: Table shows post-hoc results for the pattern run length using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

Source	SS	DF	F	p-unc
condition	15.053000	1	0.983484	3.217180e-01
pattern_type	2.582487e+06	1	2689.220036	2.751183e-208
Residual	9719.181137	635		

Table 26: Table shows ancova results for the pattern type. Condition represents the 4 samplers + human single results. Pattern_type represents the two examples of a pattern Ni+1, Ni+2, Ni+3 versus Ni+1, Ni+2, Ni+3-1 categories of patterns: Short, medium and long range jumps.

A	B	mean(A)	mean(B)	diff	se	tail	T	df	pval	hedges
direct	hmc	3.604	2.636	0.967	0.061	two-sided	11.301	1998.297	0.001000	0.506
direct	**human**	**3.604**	**5.444**	**-1.840**	**0.224**	**two-sided**	**-5.799**	**289.431**	**0.001000**	**-0.396**
direct	mc3	3.604	3.732	-0.129	0.078	two-sided	-1.172	2082.230	0.499998	-0.050
direct	mcmc	3.604	4.774	-1.171	0.088	two-sided	-9.422	2510.773	0.001000	-0.369
hmc	**human**	**2.636**	**5.444**	**-2.807**	**0.223**	**two-sided**	**-8.910**	**281.471**	**0.001000**	**-0.621**
hmc	mc3	2.636	3.732	-1.096	0.073	two-sided	-10.629	1821.129	0.001000	-0.474
hmc	mcmc	2.636	4.774	-2.138	0.084	two-sided	-18.048	2252.875	0.001000	-0.752
human	**mc3**	**5.444**	**3.732**	**1.711**	**0.228**	**two-sided**	**5.307**	**308.427**	**0.001000**	**0.362**
human	**mcmc**	**5.444**	**4.774**	**0.669**	**0.232**	**two-sided**	**2.042**	**328.707**	**0.246586**	**0.135**
mc3	mcmc	3.732	4.774	-1.042	0.097	two-sided	-7.612	2699.500	0.001000	-0.297

Table 27: Table shows post-hoc results for the pattern type using a pairwise gameshowll analysis. A/ B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler

Source	SS	DF	F	p-unc
condition	4.179201e+05	4	108.798144	2.073125e-67
pattern_jump	2.582487e+06	1	2689.220036	2.751183e-208
Residual	5.022426e+05	523		

Table 28: Table shows results for the pattern jump. Condition represents the 4 samplers + human single results. Pattern_jump represents the three categories of patterns: Short, medium and long range jumps.

A	**B**	**mean(A)**	**mean(B)**	**diff**	**se**	**tail**	**T**	**df**	**pval**	**hedges**
direct	hmc	129.15	66.29	62.86	5.25	two-sided	8.46	236.32	0.00	0.98
direct	**human**	**129.15**	**134.71**	**-5.56**	**18.75**	**two-sided**	**-0.21**	**22.59**	**0.90**	**-0.05**
direct	mc3	129.15	128.44	0.71	7.07	two-sided	0.07	289.99	0.90	0.01
direct	mcmc	129.15	169.79	-40.64	7.87	two-sided	-3.65	117.92	0.00	-0.56
hmc	**human**	**66.29**	**134.71**	**-68.42**	**18.37**	**two-sided**	**-2.63**	**20.82**	**0.08**	**-0.61**
hmc	mc3	66.29	128.44	-62.15	5.99	two-sided	-7.34	214.47	0.00	-0.84
hmc	mcmc	66.29	169.79	-103.50	6.91	two-sided	-10.58	76.48	0.00	-1.63
human	**mc3**	**134.71**	**128.44**	**6.27**	**18.97**	**two-sided**	**0.23**	**23.66**	**0.90**	**0.05**
human	**mcmc**	**134.71**	**169.79**	**-35.08**	**19.28**	**two-sided**	**-1.29**	**25.14**	**0.67**	**-0.32**
mc3	mcmc	128.44	169.79	-41.35	8.38	two-sided	-3.49	139.70	0.00	-0.54

Table 29: Table shows post-hoc results for the pattern jump using a pairwise gameshowll analysis. A B shows the 5 model comparisons. Bold values show values of interest and coloured value identify statistical non significance between human and sampler