Sampling in Human Cognition

by

Edward Vul

Submitted to the Department of Brain and Cognitive Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author .....
Department of Brain and Cognitive Science
May 03, 2010

Certified by .....
Nancy G. Kanwisher
Walter A. Rosenblith Professor of Neuroscienc
Thesis Supervisor

Accepted by .....
Earl Miller, Picower Professor of Neuroscience
Chairman, Department Committee on Graduate Theses

Sampling in Human Cognition

by

Edward Vul

Submitted to the Department of Brain and Cognitive Science
on May 03, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

Bayesian Decision Theory describes optimal methods for combining sparse, noisy data with prior knowledge to build models of an uncertain world and to use those models to plan actions and make novel decisions. Bayesian computational models predict human behavior in cognitive domains ranging from perception to motor control and language. However this success has highlighted long-standing challenges for this view that have posed a barrier to bridging the computational and process levels of cognition. First, the computations required for exact Bayesian inference are incommensurate with the limited resources available to cognition (e.g., computational speed; and memory). Second, Bayesian models describe computations but not the processes that carry out these computations and fail under cognitive load or deficits. I suggest a resolution to both challenges: The mind approximates Bayesian inference by sampling. Experiments across a wide range of cognition demonstrate Monte-Carlo-like behavior by human observers; moreover, new modeling approaches demonstrate how specific Monte Carlo algorithms can describe previously elusive cognitive phenomena. Using sampling algorithms as a process model yields methods for jointly modeling the computational and process levels, sheds light on new and old cognitive phenomena, and holds much promise for future research.

Thesis Supervisor: Nancy G. Kanwisher
Title: Walter A. Rosenblith Professor of Neuroscienc

Acknowledgments

This would not have been possible without implicit and explicit help from a large number of people.

Most importantly, I need to thank Nancy Kanwisher and Josh Tenenbaum – their patience, insights, and guidance are remarkable, and I can only hope that some day I will have a fraction of their scientific clarity and supervisory skill. Thanks also to the other committee members: Ted Adelson and Larry Maloney with whom every conversation has taught me something new.

As much as I learned from the faculty, I probably learned as much from fellow graduate students and post-docs. Thanks so much for engaging, thoughtful, and educational conversations with Mike Frank, Noah Goodman, Tim Brady, Steve Piantadosi, Talia Konkle, Vikash Mansinghka, and Danny Dilks. Thanks also to their social support, both in debauchery (Dilks), and athleticism (the BCS triathlon team – Mike, Noah, Talia, and Barbara).

Thanks also to co-authors on the articles that comprise this thesis (who were not already mentioned): Tom Griffiths, Hal Pashler, Anina Rich, and Deborah Hanus. Particularly, thanks to Deborah Hanus, who working as a UROP with me did a large proportion of the work that went into this document.

Of course – thanks to my parents, who started preparing me for this dissertation long before I knew how to read or write.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

Chapter 1

Sampling in Human Cognition

David Marr outlined three interconnected levels at which cognition may be described: *Computation* – what information is used to solve a problem and how is this information combined? *Algorithm/Process* – how is information represented and what procedures are used to combine the representations? *Implementation* – how are these representations and procedures implemented in the brain?

Bayesian inference and *decision theory* describe theoretically optimal computations for combining different sources of uncertain information to build structured models of the world and for using these models to plan actions and make decisions in novel situations. In recent years, this framework has become a popular and successful tool for describing the computations people must carry out to accomplish perceptual (Knill & Richards, 1996), motor (Maloney, Trommershauser, & Landy, 2007), memory (Anderson & Milson, 1989), and cognitive (Chater & Manning, 2006; McKenzie, 1994; Griffiths & Tenenbaum, 2005; Goodman, Tenenbaum, Feldman, & Griffiths, 2008) tasks both in the lab and in the real world; thus supporting the claim the Bayesian inference provides a promising description of *rational models* (Anderson, 1990) of cognition at the computational level. However, as Bayesian rational analysis gains ground as a computational description, several salient challenges have hampered progress at the purely computational level indicating that important constraints at the process level must be taken into account to accurately describe human cognition.

1.0.1 Challenges to a purely computational view

First, exact Bayesian calculations are practically (and often theoretically) impossible in Bayesian statistics and machine learning because computing the exact Bayesian answer often requires evaluating and integrating over innumerably large hypothesis spaces. This is true of even small-scale inferences in artificial problems (for instance – possible parses in a probabilistic context-free grammar (Charniak, 1995)), thus limiting the use of exact Bayesian inference to a small set of problems with analytically tractable posterior distributions or ones with only a small hypothesis space. Thus, applications of Bayesian inference by statisticians and computer scientists have relied on approximate inference methods. Since exact Bayesian inference is usually intractable even for small artificial problems, this must be even more true for the large-scale sorts of real-world problems that the human mind faces every day. How can the brain do Bayesian inference at a real-world scale?

Second, there is an additional practical challenge to implementing approximate statistical inference in humans: cognitive limitations. Human cognition is limited in memory (Wixted & Ebbesen, 1991; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), processing speed (Shepard & Metzler, 1971; Welford, 1952), and attention (Pashler, 1984; Broadbent, 1958; Treisman & Gelade, 1980; James, 1890), while people are faced with split-second decisions. Adequate approximate inference methods in statistics and machine learning rely on millions of complex calculations on dedicated computing clusters resulting in days of computation (Robert & Casella, 2004). What procedures can people use to approximate statistical inferences in real-world decisions within a fraction of a second, despite their limited cognitive resources?

Third, although across many domains people seem to be Bayesian on the average over many trials or subjects, individuals on individual trials are often not optimal. Goodman et al. (2008) showed that optimal average Bayesian rule-learning behavior emerges from aggregating over many subjects each of which learns just one rule. Similarly, Griffiths and Tenenbaum (2006) demonstrated that on average, people know the distribution of quantities of the world, but individual responses seemed to

reflect knowledge of only a small set of world quantities (Mozer, Pashler, & Homaei, 2008). What cognitive processes could produce optimal behavior on the average of many suboptimal trials?

Fourth, the characteristic dynamics of cognition highlight the need for a process-level description: People forget what they have learned (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), they overweight initial training (Deese & Kaufman, 1957), they solve problems (Vul & Pashler, 2007) and rotate mental images (Shepard & Metzler, 1971) slowly, and they stochastically switch between interpretations when exposed to ambiguous stimuli (Alais & Blake, 2005). Such dynamics of human cognition are outside the scope of purely computational analysis, and require a process-level description. [1]

1.1 The sampling hypothesis

We suggest a resolution to all of these challenges: The mind approximates Bayesian inference by *sampling*.

Sampling algorithms represent probabilistic beliefs by considering small sets of hypotheses randomly selected with frequency proportional to their probability. Sampling can approximate probability distributions over large hypothesis spaces despite limited resources. Sampling predicts optimal behavior on average and deviations from optimality on individual decisions. Specific sampling algorithms have characteristic dynamics that may help explain the dynamics of human cognition. Altogether, sampling algorithms are formal process-level descriptions of Bayesian inference that may help bridge the gap between ideal-observer analyses and known resource constraints

[1] It must be noted that for any pattern of behavior, it may be possible to construct a particular set of prior beliefs and utilities that would yield this pattern of behavior as the globally optimal solution.
For instance, to describe the seemingly suboptimal human behavior of slow learning and quick forgetting of, say, verbal facts (Cepeda et al., 2008); one might postulate that people have the following priors:
(a) When people are trying to teach me words, they are likely to be wrong or lying a lot of the time
(b) Whatever words that I am taught, are likely to change meaning, or be dropped from the lexicon, quite quickly.
These two assumptions will yield slow learning and rapid forgetting for an ideal computational model. However, explanations of behavior with this structure seem to overlook more general properties of cognition and are only narrowly applicable.

from cognitive psychology.

To elucidate the sampling hypothesis, it should be contrasted with a number of alternate theories of cognitive representations. The most salient accounts of cognitive representation are *Boolean point estimates* and *probability distributions*.

1.1.1 Boolean-valued point estimates

Classical accounts of neural and psychological representation presume that beliefs are represented as noisy, Boolean-valued point-estimates. Boolean-valued belief representations contain single estimates of a belief: In choices from multiple discrete options, one or more options may be deemed true, and the others false. An object either belongs to a category, or it does not. A signal has either passed threshold, or it has not. In choices along continuously valued dimensions (e.g., brightness), all-or-none representations take the form of point-estimates (e.g., 11.3 Trolands). Although the content of a point-estimate is continuous (11.3), its truth value is Boolean. Such Boolean accounts of mental representation have been postulated for signal detection (point estimates corrupted by noise; e.g., Green & Swets, 1966), memory (memory traces as point estimates; e.g., Kinchla & Smyzer, 1967), concepts and knowledge (as logical rules and Boolean valued propositions; e.g., Bruner, Goodnow, & Austin, 1956).

Such Boolean-valued belief representations fail to represent uncertainty, and as such, cannot support the Bayesian probabilistic computations that describe human behavior in these same domains: signal detection (Whitely & Sahani, 2008), memory (Steyvers, Griffiths, & Dennis, 2006), categorization (Tenenbaum, 1999), and knowledge (Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008; Vul & Pashler, 2008).

1.1.2 Full probability distributions

The opposite extreme which may be supported by a strictly computational account of Bayesian cognition would hold that cognitive representations are exact representations of probability distributions. A probability distribution may be exactly represented

in two ways. First, analytically: as a mathematical function that codifies the probability of any possible hypothesis. Obviously, the cognitive and neural plausibility of mental representations being, literally, mathematical functions, is low. Second, probability distributions may be represented as a fully enumerated weighted list: a paired list of every hypothesis along with its associated probability. Probabilistic population codes (Ma, Beck, Latham, & Pouget, 2006) effectively describe a weighted-list representation.[2].

While weighted lists may be a plausible representation for cases with fairly simple inferences, they quickly break down in the face of large-scale combinatoric problems, where the number of hypotheses grows exponentially to potentially infinite length. In these cases, a weighted list would need to be impossibly, or at least implausibly, long.

1.1.3 Sample-based representations

According to the sampling hypothesis, people represent probability distributions as sample-generating procedures, and as sets of samples that have been generated from these procedures. Inference by sampling rests on the ability to draw samples from an otherwise intractable probability distribution: to arrive at a set of hypotheses which are distributed according to the target distribution, by using a simple algorithm (such as Markov chain Monte Carlo; Robert & Casella, 2004; or particle filtering; Doucet, De Freitas, & Gordon, 2001). Samples may then be used to approximate expectations and predictions with respect to the target probability distribution, and as the number of samples grows these approximations approach the exact distributions of interest. As a physical example, consider the "plinko" machine (Figure 1-1, Galton, 1889) – this device represents a Gaussian distribution in so far as additional balls dropped in can generate additional (approximately) Gaussian samples. Representations via samples and sample-generating procedures can represent uncertainty as the variation of the set of samples (in contrast to single point-estimates). Moreover, in contrast to weighted lists, sample-based representations may be truncated at a short, finite

[2]But it should be noted that probabilistic population codes are not exact representations of probability distributions because they approximate continuous densities as a finite set of kernels.

length without introducing systematic error.

1.1.4 Theoretical considerations

Sample-based inference is typically used in machine learning and statistics to approximate Bayesian inference for a number of reasons that also make it an appealing process-model for cognitive science. First, sampling algorithms are applicable to a large range of computational models, thus affording a general inference scheme for a range of models across cognitive domains. Second, sampling algorithms scale efficiently to high-dimensional problems while minimizing the consequences of the curse of dimensionality; thus remaining plausible candidates for implementations of real-world inference. Third, sampling algorithms are a class of just-in-time algorithms that allow for a smooth tradeoff between precision, speed and computational load; thus sampling algorithms can be used under conditions of limited time and cognitive resources, while also supporting more accurate inferences when resources allow.

For our purposes, the central appeal of sampling algorithms as candidate process models are their graceful degradation with limited cognitive resources as well as their just-in-time properties. However, one would be right to ask: just how much sampling is necessary? In Bayesian statistics and machine learning, it is well-known that accurate inference requires tens or hundreds of thousands of samples, each of which is effortful to produce. We recently asked whether the same holds true when making *decisions*: How many samples are necessary to make sufficiently accurate decisions? We found that across a large range of tasks, using few sample often yields decisions that are not much worse than those based on more precise inferences (Vul, Goodman, Griffiths, & Tenenbaum, n.d.). Moreover, on the assumption that sampling is a time-consuming process, we found that using just one sample for decisions often maximizes expected reward: making quick, suboptimal decisions is often the globally optimal policy (Chapter 2).

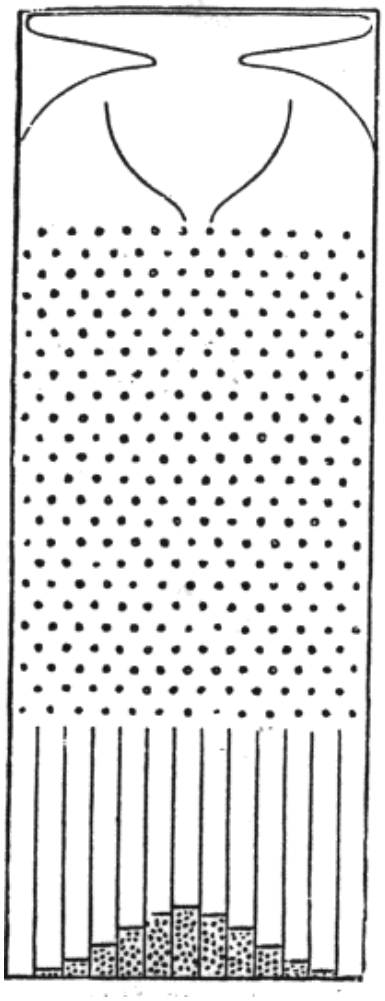

Figure 1-1: A physical example of a sampling-function representation of a probability distribution. The plinko machine (quincunx from Galton, 1889) is a device that approximately represents a Gaussian probability distribution. A ball is dropped in the top, and passes through a series of interlaced pegs – at each layer, there is a 50% chance of the ball bouncing left or right. After passing through many such layers, the final distribution of positions is Gaussian (given the central limit theorem). Thus, this box is a physical instantiation of a sample-generating process.

1.2 Relationship between sampling and classical theories

Although the sampling hypothesis is a novel process-level description that can connect computational Bayesian models, it is closely related to several classical laws and theories of cognition.

1.2.1 Probability matching

Probability matching (Herrnstein, 1961; Vulkan, 2000) refers to the characteristic behavior of people and animals when faced with a risky choice among several alternatives. In a typical probability matching experiment, an observer is faced with a choice between two options, one rewarded with probability p, the other with probability 1-p. The optimal choice is to *maximize* and always choose the option with the greater probability of reward; however, instead, people choose the two alternatives with frequency proportional to the probability of reward; thus *matching* the reward probability. Probability matching can be described as sampling from prior experience: randomly selecting previously experienced outcomes and choosing the option that was rewarded most often in the set of sampled outcomes. When only one previous trial is considered, this procedure yields probability matching, and if more trials are considered then behavior will vary between probability matching and maximizing. However, the sampling hypothesis is not restricted to previously experienced outcomes: hypotheses may be sampled not only from direct experience but also from internal beliefs inferred indirectly from observed data. Thus the sampling hypothesis is a generalization of classical probability matching, by providing predictions about cognition and behavior in cases of pure reasoning, rather than prediction from direct prior experience.

1.2.2 Generalized Luce choice axiom and soft-max link functions

Luce (1959) described a more general relationship between representations and behavior. He argued that people make responses in proportion to the strength of some internal representations:

$$P_R(a) = v(a) / \sum_b v(b), \tag{1.1}$$

where $P_R(a)$ is the probability of responding with option a, while $v(x)$ is the value of the internal representation. The soft-max generalization of the Luce choice axiom postulates that $v(x) = p(x)^L$ so that the probability of choosing an alternative is proportional to the posterior probability of that alternative raised to an exponent L. This formulation provides two interesting generalizations of simple probability matching: first, the exponent L yields a smooth gradient between probability matching and maximizing. Second, by generalizing the choice formulation to apply to any quantity of interest – not just probability of reward – this formulation allows for a link between decisions and cognitive representations across domains. Bayesian models of cognition typically rely on this soft-max link function to interface between model predictions and behavior (e.g., Frank, Goodman, & Tenenbaum, 2009; Goodman et al., 2008).

Sample-based approximations ground both of these applications of the generalized Luce choice axiom in a single process model. First, the exponent L is a proxy for the number of samples used in a decision. The exact relationship between the Luce choice exponent (L) and the number of samples (k) used to make a binary decision is analytically intractable, and varies as a function of the underlying probabilities. In all cases, however, as k increases, L increases monotonically [3]. Second, sample-

[3]The relationship between L and k is non-linear and variable: This function can be described as piecewise linear, with a steeper slope for small k and a shallower slope for larger k. For Bernoulli trials, when p is near 0.5 this transition happens around $k = 100$. As p increases, the slopes of the linear components increase, and the transition between the shallower and steeper slopes happens at smaller k values.

based inference on any problem will yield a soft-max relationship between posterior probabilities and response probabilities, thus justifying (and explaining the common need to use) soft-max link functions to connect ideal Bayesian observers to human behavior.

1.2.3 Point-estimates, noise, and drift-diffusion models

Point-estimate based representations explain variation across trials and individuals as noise corrupting point-estimates. Because the structure of the noise defines a probability distribution over possible states, the variation in responses across trials predicted by sampling and predicted by the noise accounts align under these special circumstances. Crucially, however, as I will describe below, predictions of these two accounts diverge when considering the relationship between the errors contained in different guesses.

A specific case of the noisy point-estimate account the drift-diffusion model (Ratcliff, 1978; Gold & Shadlen, 2000) allows for quantitative assessment of speed-accuracy tradeoffs on the tacit assumption that people aggregate noisy samples of input data until they reach some decision criterion. These cases may be construed as obtaining "samples" from the external world when used to account for perceptual decisions (Gold & Shadlen, 2000), but when applied to cognitive decisions, such as memory retrieval (Ratcliff, 1978), the samples must be internally generated. In these later cases, the drift-diffusion models that are superficially isomorphic to sample-based inference about internal beliefs.

Thus, the sampling hypothesis unifies internal noise, drift diffusion models, and soft-max probability matching behavior under one general framework that describes how people can approximate optimal Bayesian inference in situations without direct prior experience about the task at hand and must make decisions and inferences based solely on pure reasoning.

1.3 People seem to sample basic monte carlo

Critically, representation via point-estimates, probability distributions, or samples, make different predictions about the information contained in errors and response variability. According to the noisy Boolean point-estimate hypothesis, variation across responses arises from the accumulation of noise on internal point-estimates – thus, errors will be cumulative and correlated. If internal representations are complete probability-distributions, variation in responses can only reflect variation in inputs or utility functions. In contrast, the sampling hypothesis predicts that variation in responses arises from sampling under uncertainty and thus multiple responses will contain independent error, improve estimates when averaged, and will reflect the posterior distribution of beliefs. Thus, although using only one or a few samples for decisions will not result in substantially worse performance, it will produce characteristic response variation which we can look for in human behavior.

Although sampling – resulting in probability-matching to internal beliefs – predicts similar behavior to noisy point-estimate based representations, there is a crucial distinction in the relationship between multiple guesses. According to internal noise models, a single point-estimate is corrupted by noise, and this noise will therefore be shared by a number of guesses based on the same point estimate. In contrast, sampling based models might predict the same distribution of errors as noise based models on the first guess; however, crucially they predict independence between multiple guesses, based on the same stimulus information. This prediction has now been thoroughly tested in the case of selective visual attention: a domain where such dependencies between guesses can be precisely teased apart.

Vul, Hanus, and Kanwisher (2009) asked subjects to report one letter from a circular array cued by a line. In these situations, subjects often report nearby items instead of the target. Vul et al. asked if such errors reflect internal noise or sampling under uncertainty. Subjects made multiple guesses about the identity of the cued letter and the researchers investigated the spatial dependency between guesses. If errors in this task reflect noise corrupting the position of the cue, then there would

be a correlation in errors between the two guesses: if the first guess contained an error clockwise from the cue, the second guess should as well. However, if these errors due to sampling given uncertainty about the spatial co-occurrence of cue and target, then the errors should be independent. The authors found that the first and second guess were independent and identically distributed – this is characteristic of independent samples from a probability distribution describing uncertainty about the spatial co-occurrence of letters with the cue. This result was replicated for the case of temporal uncertainty when subjects selected items from a rapid serial visual presentation (RSVP) (Vul et al., 2009). Together, these results indicate that when subjects are asked to select items under spatiotemporal uncertainty, subjects make guesses by independently sampling alternatives from a probability distribution over space and time (Chapter 3).

This claim may be further tested and extended by asking subjects to report two orthogonal features of the target object to assess whether illusory conjunctions (Treisman & Schmidt, 1982), or misbinding errors, also arise from sampling under spatiotemporal uncertainty. Vul and Rich (in press) presented subjects with arrays and RSVP streams of colored letters and asked subjects to report both the color and the letter. Given the permuted arrangement of colors and letters, these two dimensions yielded two independent estimates of the reported spatial positions. Again, the correlation between these reports could be used to evaluate the independence of the two guesses. In this case as well, errors were independent indicating that different features are independently sampled and that illusory conjunctions and binding errors arise from spatiotemporal uncertainty (Chapter 4).

One consequence of the independent error arising from sampling under uncertainty is a prediction of a wisdom of crowds (Suroweicki, 2004) within one individual. Galton (1907) demonstrated that averaging guesses from multiple individuals yields a more accurate answer than can be obtained from one individual alone because the independent error across individuals averages out to yield a more accurate estimate. If multiple guesses from one individual are independent samples, they should also contain independent error, and then the average of multiple guesses from one individ-

ual should also yield a similar "wisdom of crowds" benefit, where the crowd is within one individual. Vul and Pashler (2008) tested this prediction in the domain of world knowledge by asking subjects to guess about world trivia (e.g., what proportion of the worlds airports are in the United States?). After subjects made one guess for each question, they were asked to provide a second guess for each. The average of two guesses was more accurate than either guess alone, indicating that the two guesses contained some independent error (despite the fact that subjects were motivated to provide their best answer on the first guess). This independence would arise if people made guesses by sampling from an internal probability distribution over estimates (Chapter 5).

The above cases verify the predictions of the sampling account in cases where the implicit computational model is not specified, but does the sampling hypothesis yield additional predictive power in cases with a concrete computational model? Goodman et al. (2008) showed that the average frequency with which subjects classify transfer items as positive instances fits almost perfectly with the probabilistic predictions of a Bayesian rule-learning model. The model considers all possible classification rules, computes a posterior probability for each rule given the training data, and then computes the probability that any item belongs to the category by averaging the decisions of all possible rules weighted by their posterior probabilities. Is this fully Bayesian inference what individual subjects do on any one trial? Not in this task. Goodman et al. (2008) analyzed the generalization patterns of individual subjects reported by (Nosofsky, Palmeri, & McKinley, 1994) and found that response patterns across seven test exemplars were only poorly predicted by the Bayesian ideal. Rather than averaging over all rules, these generalization patterns were instead consistent with each participant classifying test items using only one or a few rules; while which rules are considered varies across observers according to the appropriate posterior probabilities. Thus, it seems that individual human learners are somehow drawing one or a few samples from the posterior distribution over possible rules, and behavior that is consistent with integrating over the full posterior distribution emerges only in the average over many learners. Similar sampling-based generalization behavior

has been found in word learning (Xu & Tenenbaum, 2007) and causal learning tasks (Sobel, Tenenbaum, & Gopnik, 2004), in both adults and children.

1.4 Specific sampling algorithms for specific tasks

Although predictions of sample-based inference are confirmed across a number of domains, the fact remains that producing a sample from the appropriate posterior distribution is no trivial task. In computer science and statistics there are many algorithms available for doing Monte Carlo inference. Simple sample-generating algorithms, like rejection sampling, tend to be slow, inefficient, and computationally expensive. In practice, different sampling algorithms are chosen for particular problems where they may be most appropriate. Therefore, while "sampling" may capture some cognitive phenomena at a coarse grain, the exact sampling algorithms used may vary across domains, and may provide more accurate descriptions of specific behavioral phenomena and the dynamics of cognition.

Most real-world domains offer only small amounts of training data which must then support a number of future inferences and generalizations. Shi, Griffiths, Feldman, and Sanborn (in pressa) showed that in such domains, exemplar models (Medin & Shaffer, 1978) using only a few examples can support Bayesian inference as an importance sampler (Ripley, 1987). This can be achieved using an intuitive psychological process of storing just a small set of exemplars and evaluating the posterior distribution by weighting those samples by their probability. Shi et al. (in pressa) argued that such an importance sampler accounts for typicality effects in speech perception (Liberman, KS, Hoffman, & Griffith, 1957), generalization gradients in category learning (Shepard, 1987), optimal estimation in everyday predictions (Griffiths & Tenenbaum, 2006), and reconstructive memory (Huttenlocher, Hedges, & Vevea, 2000).

In domains where inference must be carried out online as data are coming in, such as sentence processing (Levy, Reali, & Griffiths, 2009), object tracking (Vul, Frank, Alvarez, & Tenenbaum, 2010), or change-point detection (Brown & Steyvers, 2008),

particle filtering is a natural algorithm for doing this online inference. Particle filters track a sampled subset of hypotheses as they unfold over time; at each point when additional data are observed, the current set of hypothesized states are weighted based on their consistency with the new data, and resampled accordingly – as a consequence, this inference algorithm produces a bias against initially implausible hypotheses. Levy et al. (2009) showed how this bias can account for garden-path effects in sentence processing: when the start of the sentence suggests one interpretation of the data, but the end of the sentence is disambiguated in favor of a less likely parse, the difficulty of the resampling and update step is amplified. Similar arguments have been used to explain individual differences in change detection (Brown & Steyvers, 2008), and performance while tracking objects (Vul et al., 2010).

In some real-world and laboratory tasks, the observer sees all the relevant data and must make sense of it over a period of time. For instance, when looking at at 2D projection of a wireframe cube (Necker, 1832), observers are provided with all of the relevant data at once, but must then come up with a consistent interpretation of the data. In cases where two equally likely interpretations of the stimulus are available, the perceived interpretation changes stochastically over time, jumping between two modal interpretations. Sundareswara and Schrater (2007) demonstrated that the dynamics of such rivalry in the case of a Necker cube arises naturally from approximate inference via Markov Chain Monte Carlo (MCMC; Robert & Casella, 2004). Gershman, Vul, and Tenenbaum (2010) elaborated on this argument by showing that MCMC in a coupled markov random field – like those typically used as computational models of low-level vision – not only produces bistability and binocular rivalry, but also produces the characteristic traveling wave dynamics of rivalry transitions (Gershman et al., 2010; Wilson, Blake, & Lee, 2001).

Specific sampling algorithms yield concrete predictions about online processing effects which have been inaccessible to strictly computational accounts. The dynamics of online sampling algorithms can predict learning effects, online biasing effects, as well as the specific dynamics of decision making and belief formation.

1.5 Conclusion

We started with a set of challenging questions for an ideal Bayesian description of the computational level of human cognition: How can people approximate ideal statistical inference despite their limited cognitive resources? How can we account for the dynamics of human cognition along with the associated errors and variability of human decision-making? Across a range of cognitive behaviors, sampling-based approximate inference algorithms provide an account of the process-level dynamics of human cognition as well as variation in responses.

Chapter 2

One and Done? Optimal Decisions From Very Few Samples

2.1 Thesis framing

The sampling hypothesis suggests that people approximate ideal Bayesian computations by sampling, thus allowing themselves to make near-optimal decisions in large real-world problems under time constraints and with limited cognitive resources. However, sampling itself is often a difficult procedure, and to approximate exact Bayesian inference many samples are required. In this chapter, I ask how many samples are necessary to make a *decision* and demonstrate that decisions based on few samples are quite close to optimal, and may even be globally optimal themselves, when factoring in the cost of making slow decisions.

This chapter is

(a) under review at *Psychological Review* and

(b) a much shorter version has been published as: (Vul et al., n.d.).

Abstract

In many learning or inference tasks human behavior approximates that of a Bayesian ideal observer, suggesting that, at some level, cognition can be described as Bayesian inference. However, a number of findings have highlighted an intriguing mismatch between human behavior and standard assumptions about optimality: people often

appear to make decisions based on just one or a few samples from the appropriate posterior probability distribution, rather than using the full posterior distribution. Although sampling-based approximations are a common way to implement Bayesian inference, the very limited numbers of samples often used by humans seem insufficient to approximate the required probability distributions very accurately. Here we consider this discrepancy in the broader framework of statistical decision theory, and ask: if people are making decisions based on samples but samples are costly, how many samples should people use to optimize their total expected or worst-case reward over a large number of decisions? We find that under reasonable assumptions about the time costs of sampling, making many quick but locally suboptimal decisions based on very few samples may be the globally optimal strategy over long periods. These results help to reconcile a large body of work showing sampling-based or probability-matching behavior with the hypothesis that human cognition can be understood in Bayesian terms, and suggest promising future directions for studies of resource-constrained cognition.

2.2 Introduction

Across a wide range of learning, inference and decision tasks, it has become increasingly common to analyze human behavior through the lens of optimal Bayesian models (in perception: Knill & Richards, 1996; motor action: Maloney et al., 2007; language: Chater & Manning, 2006; decision making: McKenzie, 1994; causal judgments: Griffiths & Tenenbaum, 2005; and concept learning: Goodman et al., 2008). However, despite the many observed parallels, the argument for understanding human cognition as a form of Bayesian inference remains far from complete. This paper addresses two challenges. First, while human behavior often appears to be optimal when averaged over multiple trials and subjects, it may not look that way within individual subjects or trials. There will always be variance across these dimensions in any behavioral experiment, but the micro-level variation observed in many studies comparing human behavior to Bayesian models is not simply random noise around the model predictions. What kind of online processing is going on inside individual subjects' minds that can appear so different at the local scale but approximate optimal behavior when averaged over many subjects or many trials? Second, while ideal Bayesian computations are algorithmically straightforward in most small laboratory tasks, they are intractable for large-scale problems such as those that people face in the real world, or those that most Bayesian machine learning and artificial intelligence systems focus on. If human cognition is to be understood as a kind of Bayesian inference, we need an account of how the mind rapidly and effectively approximates these intractable calculations in the course of online processing.

Here we argue that both of these challenges can be resolved by viewing cognitive processing in terms of stochastic sampling algorithms for approximate Bayesian inference, and analyzing the cost-benefit tradeoff underlying the question of "How much to think?". Standard analyses of decision-making as Bayesian inference assume that people should seek to maximize the expected utility (or minimize the expected cost)

of their actions, relative to their posterior distribution over hypotheses. We show that in many settings, this ideal behavior can be approximated by an agent who considers only a small number of samples from the Bayesian posterior, and that the time cost to obtain more than a few samples outweighs the expected gain in decision accuracy they would provide. Hence human cognition may approximate globally optimal behavior by making a sequence of noisy, locally suboptimal decisions – much as we see when we look closely at individual experimental subjects and trials.

This first challenge – accounting for behavior within individual subjects and trials – is highlighted by an intriguing observation from Goodman et al. (2008) about performance in classic categorization tasks. Typically subjects learn to discriminate positive and negative exemplars of a category, and are then asked to generalize the learned rules to new transfer items. Goodman et al. (2008) showed that the average frequency with which subjects classify transfer items as positive instances fits almost perfectly with the probabilistic predictions of a Bayesian rule-learning model (Figure 2-1a). The model considers all possible logical rules for classification (expressed as disjunctions of conjunctions of Boolean features), computes a posterior probability for each rule given the training data, and then computes the probability that any item is a positive instance by averaging the decisions of all possible rules weighted by their posterior probabilities. Do individual subjects compute this same average over all possible rules in their heads on any one trial? Not in this task. Goodman et al. (2008) analyzed the generalization patterns of more than 100 individual subjects reported by (Nosofsky et al., 1994) and found that the response patterns across seven test exemplars were only poorly predicted by the Bayesian ideal, even allowing for random response noise on each trial (Figure 2-1b). Rather than averaging over all rules, these generalization patterns were consistent with each participant classifying test items using only one or a few rules; while which rules are considered varies across observers according to the appropriate posterior probabilities (Figure 2-1c). Thus, it seems that individual human learners are somehow drawing one or a few samples from the posterior distribution over a complex hypothesis space, and behavior that is consistent with integrating over the full posterior distribution emerges only in the average over many learners. Similar sampling-based generalization behavior has been found in word learning (Xu & Tenenbaum, 2007) and causal learning tasks (Sobel et al., 2004), in both adults and children.

This sampling behavior is not limited to categorization tasks but has been found in many other higher-level cognitive settings. For example, Griffiths and Tenenbaum (2006) studied people's predictions about everyday events, such as how long a cake will bake given that it has been in the oven for 45 minutes. They found a close match between the median subjects' judgments and the posterior medians of an optimal Bayesian predictor (Figure 2-2a). But the variation in judgments across subjects suggests that each individual is guessing based on only one or a small number of samples from the Bayesian posterior (c.f. Mozer et al., 2008), and the distribution of subjects' responses looks almost exactly like the Bayesian posterior distribution, rather than the optimal choice under the posterior perturbed by random response noise. Figure 2-2 shows the comparison of median human judgments with Bayesian posterior medians, along with the full quantile-quantile plots relating human and

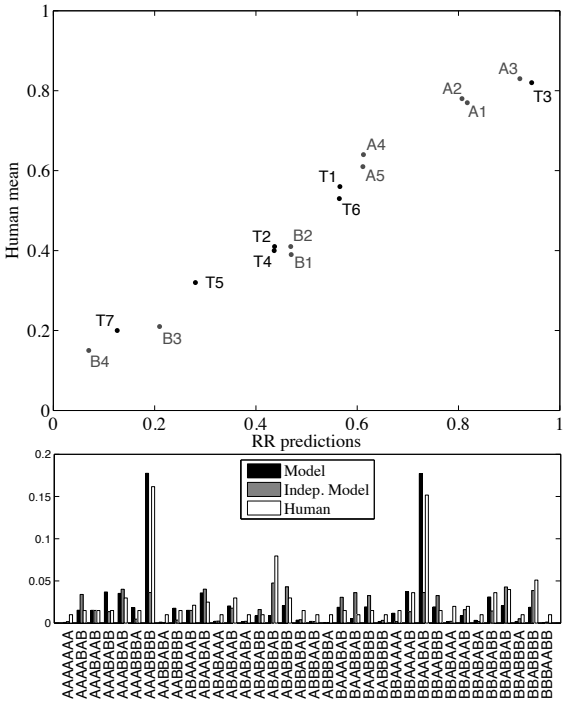

Figure 2-1: (Top) Generalization to new exemplars by subjects who learned a categorization rule is almost perfectly predicted by the ideal Bayesian model that learns a posterior over categorization rules, and then makes responses for each exemplar by considering this complete probability distribution (as shown by a very high correlation between model predictions, and predicted categorization probability; Goodman et al., 2008). (Bottom) However, the histogram of generalization patterns for seven test stimuli (white bars) does not match this ideal observer (grey bars). Generalization patterns seem to reflect a much greater correlation of beliefs from one test probe to the next than would be predicted by a model where individuals make independent judgements to test stimuli, based on the full posterior over rules. Instead, generalization patterns are consistent with individual subjects adopting one, or a few, rules in proportion to their posterior probability, and making many generalization responses accordingly (black bars; Goodman et al., 2008). Bayesian behavior emerges only on average, while individual subjects seem to reflect just a few samples from the posterior.

model predictions for seven different classes of everyday events, and an aggregate plot combining these data. While there are some deviations in specific cases – such as a tendency to produce tighter predictions than the posterior for human lifespans – the aggregate results show a close match between the two probability distributions, consistent with the idea that people are making predictions by sampling from the posterior distribution.

Further evidence that people make predictions by sampling comes from studies in which individuals must produce more than one judgment on a given task. Multiple guesses from one individual have been found to have independent errors, like independent samples from a probability distribution, when people are making estimates of esoteric quantities in the world (Vul & Pashler, 2008) or in guesses about cued visual items (Vul et al., 2009), and in illusory conjunctions in visual attention tasks (Vul & Rich, in press). More broadly, models of category learning (Sanborn & Griffiths, 2008; Sanborn, Griffiths, & Navarro, 2006; Shi, Griffiths, Feldman, & Sanborn, in pressb), change detection (Brown & Steyvers, 2008), associative learning (Daw & Courville, 2008), and language learning (Xu & Tenenbaum, 2007), have explicitly or implicitly relied on a sampling process like probability matching (Herrnstein, 1961) to link the ideal Bayesian posterior to subjects' responses, indicating that in many cases when Bayesian models predict human behavior, they do so through the assumption that people sample instead of computing the response that will maximize expected utility under the full posterior distribution.

The second challenge – that Bayesian inference is intractable – comes from the challenges that are produced in scaling probabilistic models to real-world problems. For problems involving discrete hypotheses about the processes that could have produced observed data, the computational cost of Bayesian inference increases linearly with the number of hypotheses considered. The number of hypotheses can increase in any problem that has combinatorial structure. For example, the number of causal structures relating a set of variables increases exponentially in the number of variables (with over three million possible structures for just six variables), and the number of clusterings of a set of objects increases similarly sharply (with over a hundred thousand partitions of just ten objects). In other cases, we need to work with infinite discrete hypothesis spaces (as when parsing with a recursive grammar), or continuous hypothesis spaces where there is no direct way to calculate the integrals required for Bayesian inference. The high computational cost that results from using probabilistic models has led computer scientists and statisticians to explore a variety of approximate algorithms, with exact computations being the exception rather than the rule in implementations of Bayesian inference.

Within cognitive science, critics of the Bayesian approach have seen these challenges as serious enough to question the whole program of Bayesian cognitive modeling. One group of critics (e.g., Mozer et al., 2008) has suggested that although many samples may adequately approximate Bayesian inference, behavior based on only a few samples is fundamentally inconsistent with the hypothesis that human cognition is Bayesian. Others highlight the second challenge and argue that cognition cannot be Bayesian inference because exact Bayesian calculations are computationally intractable, so the brain must rely on computationally efficient heuristics (e.g.,

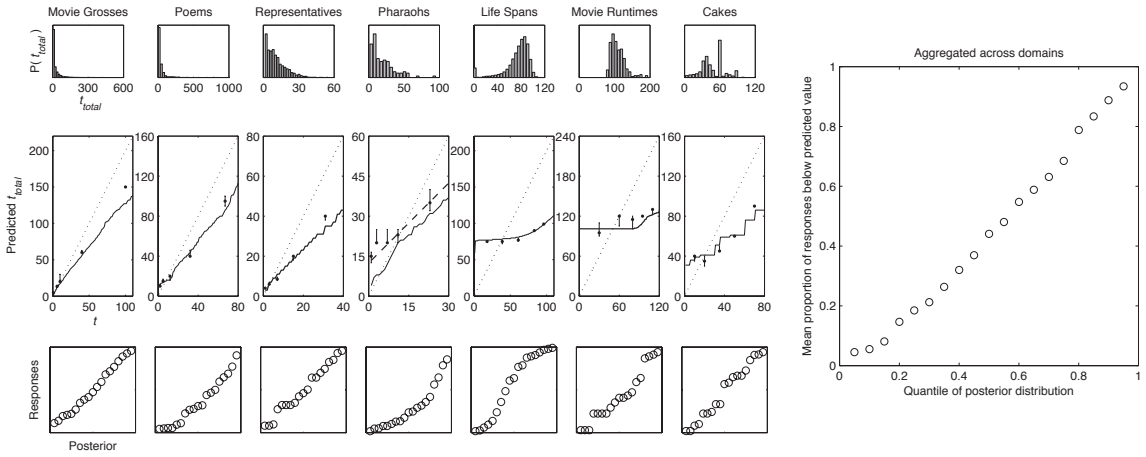

Figure 2-2: Data from Griffiths & Tenenbaum (2006) showing optimal predictions for everyday quantities. (Left, top row) The real empirical distributions of quantities across a number of domains; from left to right: movie grosses, poem lengths, time served in the US House of Representatives, the reign of Egyptian pharaohs, human life spans, movie runtimes, and the time to bake a cake. (Left, middle row) When participants are asked to predict the total quantity based on a partial observation (e.g., what is the total baking time of a cake given that it has been baking for 45 minutes?) they make predictions that appear to match the Bayesian ideal observer that knows the real-world distribution. Thus, it would appear that in all of these domains, people know and integrate over the full prior distribution of (e.g.) cake baking times when making one prediction. (Left, bottom row) However, the quantile-quantile plots comparing the distributions of human predictions with the corresponding posterior distributions reveal a different story. For each prediction, the quantiles of human response distributions were computed, and compared with the corresponding posterior distribution produced by using Bayesian inference with the appropriate prior (to produce each plot, quantiles were averaged across five predictions for each phenomenon). A match between the Bayesian posterior distribution and the distribution of people's responses corresponds to data points following along a diagonal line in these plots – where the quantiles of the two distributions are in direct correspondence. (Right) The correspondence between the posterior predictive and human responses is most pronounced when considering the quantile-quantile plot that reflects an aggregate over all seven individual quantities. Thus, people make guesses with frequency that matches the posterior probability of that answer, rather than maximizing and choosing the most likely alternative. This indicates that although participants *know* the distribution of cake baking times (as evidenced by the quantile-quantile match), they do not produce the optimal Bayesian response by integrating over this whole distribution, but instead respond based on only a small number of sampled baking times.

Gigerenzer, 2008). Addressing these challenges is thus an important step towards increasing the psychological plausibility of probabilistic models of cognition.

In this paper we will argue that acting based on a few samples can be easily reconciled with optimal Bayesian inference and may be the method by which people approximate otherwise intractable Bayesian calculations. Our argument has three central claims. First, that sampling behavior can be understood in terms of sensible sampling-based approaches to approximating intractable inference problems of the kind used in Bayesian statistics and computer science. Second, that very few samples from the Bayesian posterior are often sufficient to obtain approximate predictions that are almost as good as predictions computed using the full posterior. And third, that under conservative assumptions about how much time it might cost to produce a sample from the posterior, making predictions based on very few samples (even just one), can actually be the globally optimal strategy.

2.3 Approximating Bayesian inference by sampling

Bayesian probability theory prescribes a normative method for combining prior knowledge with observed data, and making inferences about the world. However, the claim that human cognition can be described as Bayesian inference does not imply that people are doing exact Bayesian inference. Exact Bayesian inference amounts to fully enumerating hypothesis spaces every time beliefs are updated with new data. This is computationally intractable in any large-scale application, so inference must be approximate. As noted earlier, this is the case in Bayesian artificial intelligence and statistics, and is even more relevant to solving the kinds of problems we associate with human cognition, where the real-world inferences are vastly more complex and responses are time-sensitive.

The need for approximating Bayesian inference leaves two important questions. For artificial intelligence and statistics: What kinds of approximation methods work best to approximate Bayesian inference? For cognitive science and psychology: What kinds of approximation methods does the human mind use? In the tradition of rational analysis (Anderson, 1991), or analysis of cognition at Marr's (1982) computational level, one strategy for answering the psychological question begins with good answers to the engineering question. Thus, we will explore the hypothesis that the human mind approximates Bayesian inference with some version of the algorithmic strategies that have proven best in artificial intelligence and statistics, on the grounds of computational efficiency and accuracy.

In artificial intelligence and statistics, one of the most common methods for implementing Bayesian inference is with sample-based approximations. Inference by sampling rests on the ability to draw samples from an otherwise intractable probability distribution — that is to arrive at a set of hypotheses which are distributed according to the target distribution, by using a simple algorithm (such as Markov chain Monte Carlo; Robert & Casella, 2004; or particle filtering; Doucet et al., 2001). Samples may then be used to approximate expectations and predictions with respect to the target probability distribution, and as the number of samples grows these ap-

proximations approach the exact quantities[1]. Sampling methods are typically used because they are applicable to a large range of computational models, are robust to increasing dimensionality, and degrade gracefully when computational resources limit the number of samples that can be drawn.

Computer scientists and statisticians use a wide range of sampling algorithms. Some of these algorithms have plausible cognitive interpretations, and specific algorithms have been proposed to account for aspects of human behavior (Sanborn et al., 2006; Levy et al., 2009; Brown & Steyvers, 2008; Shi, Feldman, & Griffiths, 2008). For our purposes, we need only assume that a person has the ability to draw samples from the hypothesis space according to the posterior probability distribution[2]. Thus, it is reasonable to suppose that people can approximate Bayesian inference via a sampling algorithm, and evidence that humans make decisions by sampling is not in conflict with the hypothesis that the computations they are carrying out are Bayesian.

However, using an approximation algorithm can often result in strong deviations from exact Bayesian inference. In particular, poor approximations can be produced when the number of samples is small. Recent empirical results suggest that if people are sampling from the posterior distribution, they base their decisions on very few samples (Vul & Pashler, 2008; Goodman et al., 2008; Mozer et al., 2008) – so few that any claims of convergence to the real probability distribution do not hold. Algorithms using only a few samples will have properties quite different from full Bayesian integration. This leaves us with the question: How bad are *decisions* based on few samples?

2.4 Two-alternative decisions

To address the quality of decisions based on few samples, we will consider performance of an ideal Bayesian agent (maximizing expected utility under the full posterior distribution over hypotheses) and a sample-based agent (maximizing expected utility under a small set of sampled hypotheses). We will start with the common scenario of choosing between two alternatives. Many experimental tasks in psychology are a variant of this problem: given everything observed, make a two-alternative forced-choice (2AFC) response. Moreover, real-world tasks often collapse onto such simple 2AFC decisions, for instance: we must decide whether to drive to the airport via the bridge or the tunnel, depending on which route is likely to have least traffic. Although

[1]The Monte Carlo theorem states that the expectation over a probability distribution can be approximated from samples:

$$E_{P(S)}[f(S)] \simeq \frac{1}{k}\sum_{i=1}^{k} f(S_i), \text{ when } S_i \sim P(S). \tag{2.1}$$

[2]Other authors have suggested that people sample the available data, rather than hypotheses (N. Stewart, Chater, & Brown, 2006). We focus on the more general setting of hypothesis sampling, though many of our arguments hold for data sampling as well.

this decision will be informed by prior experiences that produced intricate cognitive representations of possible traffic flow, at the moment of decision these complex representations collapse onto a prediction about a binary variable: Is it best to turn left or right?

2.4.1 Bayesian and sample-based agents

Statistical decision theory (Berger, 1985) prescribes how information and beliefs about the world and possible rewards should be combined to define a probability distribution over possible payoffs for each available action (Maloney, 2002; Kording, 2007; Yuille & Bülthoff, 1996). An agent trying to maximize payoffs over many decisions should use these normative rules to determine the expected payoff of each action, and choose the action with the greatest expected payoff[3]. Thus, the standard for decisions in statistical decision theory is to choose the action (A^*) that will maximize expected utility ($U(A;S)$) of taking an action under the posterior distribution over possible current world states (S) given prior data (D):

$$A^* = \arg\max_A \sum_S U(A;S)P(S|D). \tag{2.2}$$

To choose an action, the only property of world states we care about is the expected utility of possible actions given that state. Thus, if there are two possible actions (A_1 and A_2) and one action is "correct" (that is, there are two possible values for $U(A;S)$ and only one action for each state receives the higher value)[4] then we may collapse the state space onto a binary space: Is A_1 correct or A_2? Under this projection the posterior distribution becomes a Bernoulli distribution, where the posterior probability that A_1 is correct is p—this quantity fully parameterizes the problem, with respect to the 2AFC task. The ideal Bayesian agent who maximizes expected utility will then choose the action which is most likely to be correct (the *maximum a posteriori*, MAP, action, and will be correct p proportion of the time. (In what follows we assume p is between 0.5 and 1, without loss of generality.)

A sample-based agent samples possible world states (S_n) from the posterior distribution, uses those samples to estimate the expected utility of each action, and makes a decision based that estimate:

$$\begin{aligned} A^* = \arg\max_A \sum_{i=1}^{k} U(A;S_i) \\ S_i \sim P(S|D). \end{aligned} \tag{2.3}$$

Under the assumption that the utility has two values ("correct"/"incorrect"), the sample-based agent will thus choose the action which is most frequently correct in

[3]An agent might have other goals, e.g., maximizing the minimum possible payoff (i.e., extreme risk aversion); however, we will not consider situations in which such goals are likely.

[4]The analysis becomes more subtle when the utility structure is more complex. We return to this point in the discussion.

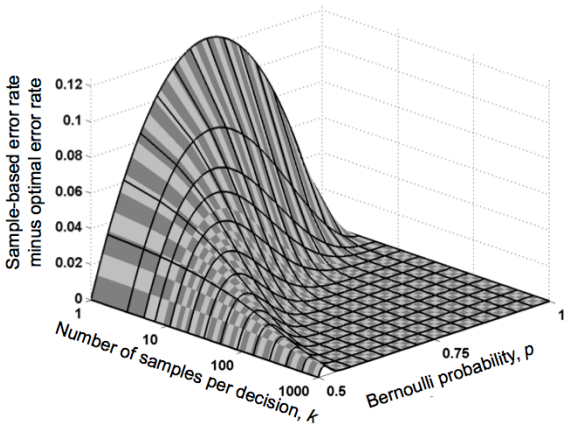

Figure 2-3: Increased error rate for the sample-based agent over the optimal agent as a function of the probability that the first action is correct and the number of samples drawn for a decision (decisions based on 0 samples not shown).

the set of sampled world states. Thus, a sample-based agent drawing k samples will choose action A_1 with probability:

$$q = 1 - \Theta_{CDF}(\lfloor \frac{k}{2} \rfloor, p, k), \tag{2.4}$$

where Θ_{CDF} is the binomial cumulative density function describing the probability that fewer than half ($\lfloor \frac{k}{2} \rfloor$) of k samples will suggest that the correct action is the best one, given that the posterior probability of the correct action is equal to p over the set of all possible samples. Thus, q is the probability that the majority of samples will point to the correct (MAP) action. Therefore, the sample-based agent will be right with probability $qp + (1-q)(1-p)$.

2.4.2 Good decisions from few samples

So, how much worse will such 2AFC decisions be if they are based on a few samples rather than an inference computed by using the full posterior distribution? Bernoulli estimated that more than 25,000 samples are required for "moral certainty" about the true probability of a two-alternative event (Stigler, 1986).[5] Although Bernoulli's calculations were based on different derivations than those which are now accepted (Stigler, 1986), it is undeniable that *inference* based on a small number of samples differs from the exact Bayesian solution and will contain greater errors, but how bad are the *decisions* based on this inference?

In Figure 2-3 we plot the difference in error rates between the sample-based and optimal agents as a function of the underlying probability (p) and number of samples (k). When p is near 0.5, there is no use in obtaining any samples (since a perfectly

[5]Bernoulli considered *moral certainty* to be at least 1000:1 odds that the true ratio will be within $\frac{1}{50}$ of the measured ratio.

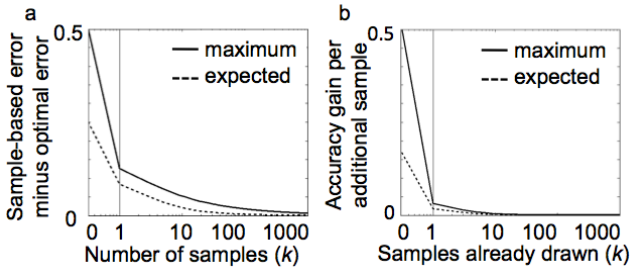

Figure 2-4: Increased error rate for the sample-based agent in 2AFC decisions marginalizing over the Bernoulli probability (assuming a uniform distribution over p). (a) The maximum and expected increase in error for the sample-based agent compared to the optimal agent as a function of number of samples (see text). (b) Expected and maximum gain in accuracy from an additional sample as a function of the number of samples already obtained.

informed decision will be as likely to be correct as a random guess). When p is 1 (or close), there is much to be gained from a single sample—since that one sample will indicate the (nearly-deterministically correct) answer; however, subsequent samples are of little use, since the first one will provide all the gain there is to be had. Most of the benefit of large numbers of samples occurs in interim probability values (around 0.7 and lower).

Since the sample-based agent does not know what the true probability p may be for a particular decision we can consider the scenarios such an agent should expect: the average scenario (expectation over p) and the worst case scenario (maximization of the loss over p). These are displayed in Figure 2-4a assuming a uniform probability distribution over p. The deviation from optimal performance decreases to negligible levels with very few samples, suggesting that the sample-based agent need not have more than a few samples to approximate ideal performance. We can go further to assess just how much is gained (in terms of decreased error rate) from an additional sample (Figure 2-4b). Again, the vast majority of accuracy is gained with the first sample, and subsequent samples do very little to improve performance.

Thus, even though few samples will not provide a very accurate estimate of p—definitely not sufficient to have "moral certainty"—they *are sufficient to choose an action*: We do not need moral certainty to act optimally.

2.4.3 How many samples for a decision?

If people do make inferences based on samples, but samples are costly, how many samples should people use before making a decision? For instance, how many possible arrangements of traffic across the city should we consider before deciding whether to turn left for the tunnel or right for the bridge? Considering one such possibility requires concerted thought and effort—it seems obvious that we should not pause at the intersection for several hours and enumerate all the possibilities. It also seems

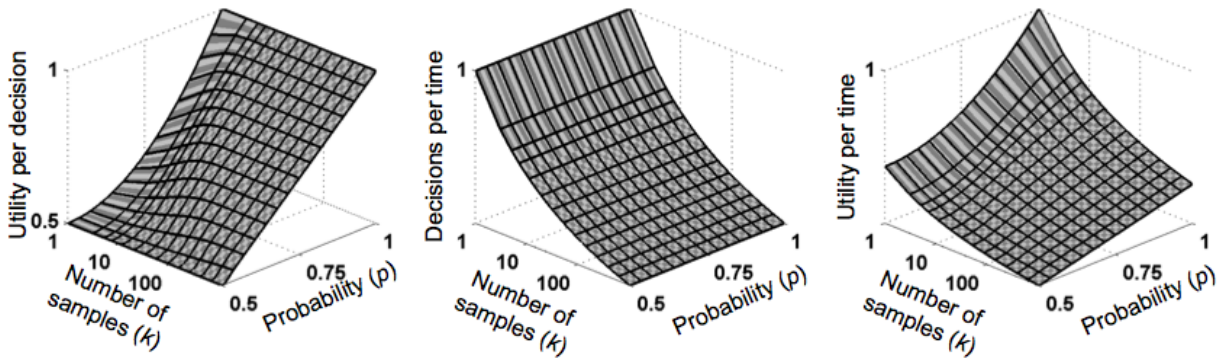

Figure 2-5: Expected utility per decision, the number of decisions that can be made per unit time, and the expected rate of return (utility per unit time) as a function of the probability that the first action is correct and the number of samples (with an example action/sample cost ratio of 232, arbitrarily chosen from one of the logarithmically spaced cost ratios we evaluated).

likely that we shouldn't just turn left or right at random without any consideration. So, how many samples should we take: how hard should we think?

Determining an optimal answer to this meta-cognitive problem requires that we specify how much a sample may "cost". To be conservative (and for the sake of simplicity), we will assume that a sample can only cost time—it takes some amount of time to conjure up an alternate outcome, predict its value, and update a decision variable.

If a given sample is free (costs 0 time), then we should take infinitely many samples, and make the best decision possible every time. If a sample costs 1 unit of time, and the *action time* (the time that it would take us to act once we have chosen to do so) is also 1 unit of time, then we should take zero samples: we should guess randomly. To make this peculiar result intuitive, lets be concrete: if we have 100 seconds, and the action time is fixed to be 1 second, then we can make 100 random decisions, which will be right 50% of the time, thus giving us an expected reward of \$50 (assuming correct choices pay \$1, and incorrect choices are not penalized). If taking a single sample to improve our decision will cost an additional second per decision, then if we take one sample per decision, each decision will take 2 seconds, and we could make at most 50 of them. It is impossible for the expected reward from this strategy to be greater than guessing randomly, since even if 100% of the decisions are correct, only \$50 will be gained. Moreover, since 100% accuracy based on one sample is extremely unlikely (this could only arise in a completely deterministic prediction task), substantially less reward should be expected. Thus, if obtaining a sample takes as long as the action, and we do not get punished for an incorrect answer, we should draw zero samples per decision and make as many random decisions as we can. More generally, we can parameterize how much a sample "costs" as the ratio between the time required to make an action and the time required to obtain one sample (action/sample ratio)—intuitively, a measure of how many samples it would take to double the time spent

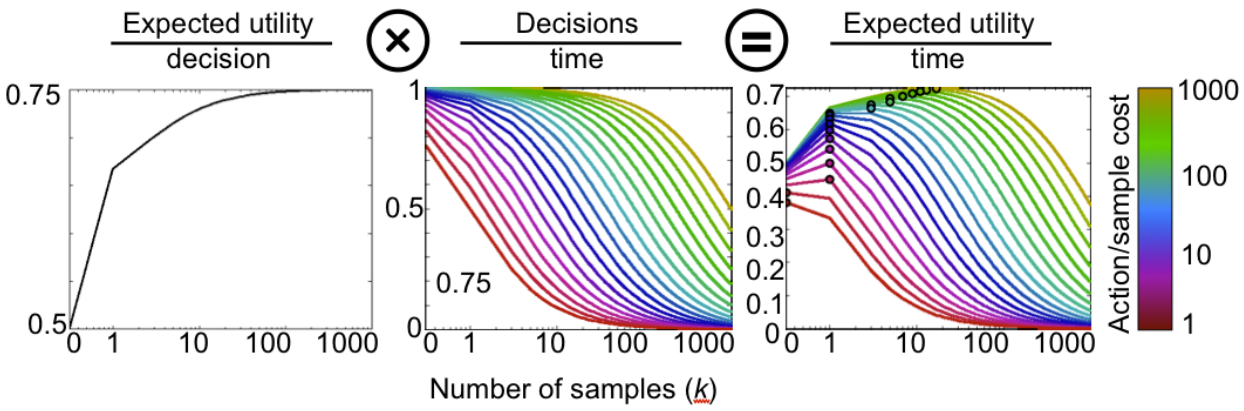

Figure 2-6: Expected utility per decision, number of decisions per unit time, and expected utility per unit time (rate of return) as a function of the number of samples and action/sample cost ratios. Action/sample cost ratios are logarithmically spaced between 1 (red) and 1000 (yellow). In the last graph the circles indicate the optimal number of samples at that action/sample cost ratio. (The utility function used for this figure contained no punishment for an incorrect choice and +1 reward for a correct choice.)

on a decision compared to making the decision using no samples.

The expected accuracy for a sample-based agent (previous section) gives us the expected utility per decision as a function of k (the number of samples) and p (the probability that the first action is correct; Figure 2-6a), and the utility function. We consider two utility functions for the 2AFC case: *no punishment* —correct: gain 1; incorrect: lose 0) and *symmetric* —correct: gain 1; incorrect: lose 1. Given one particular action/sample time ratio, we can compute the number of decisions made per unit time (Figure 2-6b). Multiplying these two functions together yields the expected utility per unit time (Figure 2-6c).

Since p is unknown to the agent, an ideal k must be chosen by taking the expectation over p. This marginalization (assuming a uniform distribution over p) for many different action/sample time ratios is displayed in Figure 2-7. It is clear that as samples become cheaper, one is best advised to take more of them—converging to the limit of infinitely many samples when the samples are free (the action/sample time ratio is infinity).

In Figure 5 we plot the optimal number of samples as a function of the action/sample time ratio. Remarkably, for ratios less than 10, one is best advised to make decisions based on only one sample if the utility function is symmetric. Moreover, with no punishment for incorrect answers, the action/sample time ratio must be 2 or greater before taking any samples becomes a prudent course of action. Thus, under a wide range of assumptions about how much it costs to think, making guesses based on very few samples (e.g., one) is the best course of action: Making many locally suboptimal decisions quickly is the globally optimal strategy.

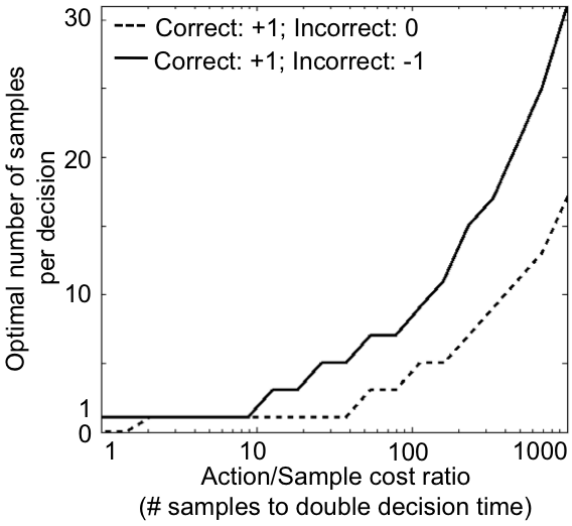

Figure 2-7: The optimal number of samples as a function of the action/sample time-cost ratio for each of two utility functions (symmetric—correct: +1, incorrect: -1; and no punishment for incorrect answers—correct: +1, incorrect: 0).

2.5 N-Alternative Decisions

So far we have only considered two alternative decisions: in such cases, no matter how high-dimensional the state of the world may be, the decision collapses onto one binary variable. It is likely that our analysis would produce different results when more than two alternatives are available (and thus, more information is required to choose among them). Therefore, we now ask the same questions of N-alternative forced choice tasks, where N is 4, 8, 16, and 32: How bad are choices among many alternatives if such decisions are based on few samples? And how many samples *should* we use when we are faced with such a decision?

On the assumption that the utility functions for such N-AFC decisions is that one and only one of the N alternatives is "correct" and the others are "incorrect", the optimal agent (who knows the multinomial distribution describing the probability that any one choice is "correct") will always choose the alternative that has the highest probability (MAP), and will be right with that probability — $\max p$; thus the performance of the optimal agent only depends on $\max p$. The sample-based agent, just as in the 2-AFC case, will choose the alternative that was "correct" under the most samples. Therefore, we show the inflation of error rates for the sample-based agent over the optimal agent as a function of the number of samples and the maximum probability of the multinomial (Figure 2-8). Just as in the 2AFC case, the optimal agent has the greatest advantage in problems of "interim" difficulty—When the maximum probability is neither too close to chance (where the sample-based agent and the optimal agent must both resort to random guessing), nor too close to certainty (when one sample will be sufficient for perfect accuracy for the sample based agent). And again, just as in the 2AFC case, the advantage of many samples

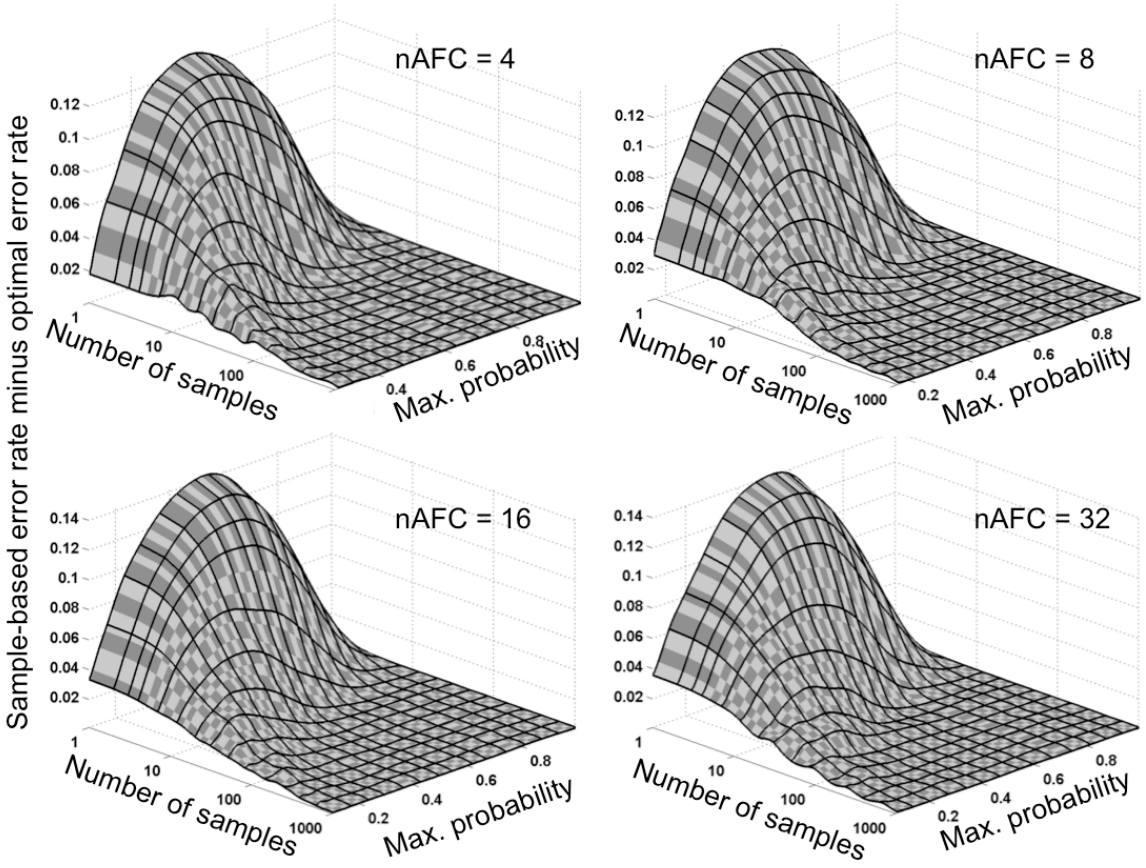

Figure 2-8: Increased error rate for the sample-based agent over the optimal agent as a function of the number of alternatives in the decision (different panels), the number of samples, and the probability of the highest-probability alternative. These figures values were produced by numerical simulation.

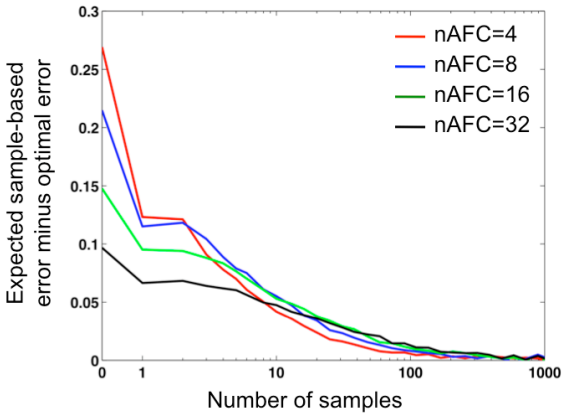

Figure 2-9: Expected increased error rate (see text for details of the marginalization) for the sample-based agent over the optimal agent as a function of the number of alternatives in the decision (different lines) and the number of samples.

decreases quickly.

Again, the relevant question is: how much worse should the sample-based agent *expect* to fair, given that probabilities are unknown. Thus, we again marginalize over possible probability distributions over alternatives, assuming a uniform prior over multinomial distributions. We marginalize over possible multinomial distributions and obtain the expected additional error for the sample-based agent over the optimal agent as a function of the number of samples (Figure 2-9). And again, just as in the 2AFC case, we see that the expected additional error decreases quickly (albeit faster for choices with fewer alternatives).

Finally, we ask: how many samples *should* the sample-based observer take when faced with a choice among many alternatives? We take the same analysis strategy as in the 2AFC case: we assume that a given sample costs time, and thus slows down the decision, and that a rational sample-based agent is trying to maximize expected rate of return. As such, we can multiply the expected utility (here we consider only the "no punishment" utility function[6]) by the number of decisions made per unit time, for each number of samples obtained. This interim calculation is shown in Figure 2-10.

From the calculation in Figure 2-10, we can then plot the optimal number of samples given a particular sample cost, for decisions with different numbers of alternatives. This is displayed in Figure 2-11. Just as in the 2AFC case, a large regime of possible sample-costs results in 1 as the optimal number of samples. However, the more alternatives there are, the faster the optimal number of samples rises as a function of decreasing sample cost; reaching an optimal calculation of as many as 75

[6]With punishment for an incorrect decision among many alternatives, it is very easy for the expected reward to be negative, rather than positive, in which case the optimal agent will try to *minimize* the number of decisions made per second — we avoid this degenerate scenario by not considering punishment.

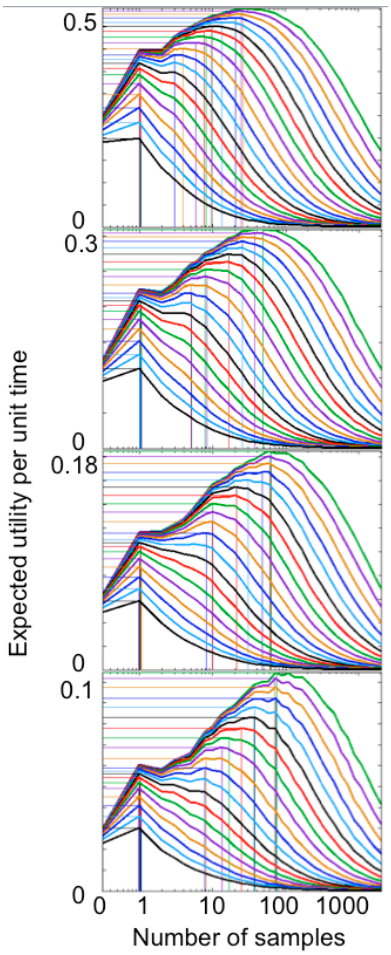

Figure 2-10: Expected rate of return for the sample-based agent as a function of number of alternatives in a decision (different panels), the number of samples used per decision, and the cost, in time, per decision (different lines).

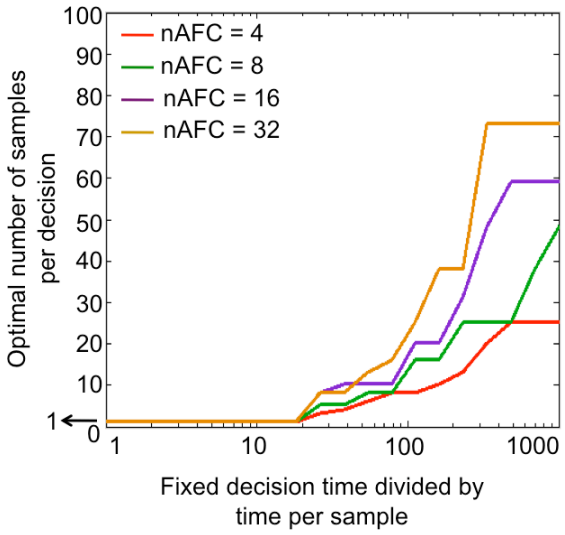

Figure 2-11: Optimal number of samples for the sample-based agent as a function of the cost, in time, of a sample (horizontal axis), and the number of alternatives in the decision being made (different lines).

samples within our tested range. Nonetheless, again, we see that in a large range of possible sample costs, making very quick, suboptimal decisions is the best policy, even when choosing among as many as 32 alternatives.

2.6 Continuous Decisions

Thus far we have shown that for choices among 2, 4, 8, 16, and 32 alternatives, a sample-based agent is often best advised to make decisions based on very few samples — thus, it should not be surprising that people are often observed to make decisions as though they are taking only a few samples in such scenarios. However, many human actions are defined over continuous variables: where to move an arm, how long to wait for a tardy friend, etc. These are all "continuous" decisions, rather than choices among discrete alternatives. These scenarios never have a single, explicitly correct answer, but rather, are often rewarded for precision—the closer to the optimal answer, the higher the reward. We will now consider actions defined by continuous variables, and again ask: How bad are decisions based on few samples, and how many samples should a sample-based agent use?

Just as in the binomial (2AFC) and multinomial (nAFC) cases, we assume that for continuous choices the "correct" answer on a given trial is drawn from the posterior probability distribution that the optimal observer has access to (and which the sample-based agent is approximating with samples). For simplicity, we will assume that here this posterior probability distribution takes the form of a Gaussian with standard deviation σ_P.

2.6.1 Making continuously-valued decisions

The best decisions for optimal and sample-based observers were straightforward for choices among a set of discrete alternatives with a utility function that classifies one alternative as correct, and all others as incorrect: choose the alternative with highest probability (or the most samples). However, when choosing along a continuous dimension, this formalization no longer makes sense. The task would be hopeless if the reward structure were a delta function — that is, if only one of infinitely many continuously varying possibilities was deemed "correct" and rewarded.

Therefore, instead of structuring rewards as a delta function, it is common practice to define a reward-function for continuous decisions that decreases as a function of distance from the correct answer. For instance, the target for archery competitions is composed of many concentric circles, and archers attempt to get an arrow as close as possible to the center, because the inner circles are worth more points. Typical reward functions for such games are different from the loss functions considered in statistics, which are commonly unbounded (for instance L2: loss that increases with the square of the distance from the target): ranging from zero for a perfect answer to infinite loss. However, in the games we consider, and arguably in the real world, the loss function drops off until it reaches some bound — if one were to miss the archery target altogether, one gets zero points, regardless of how badly the target was missed. A variant of such a utility function has been characterized mathematically as a "maximum local mass" loss function: essentially a utility function that is shaped like a gaussian probability distribution peaking at the correct answer and dropping off to zero with distance (Brainard & Freeman, 1997). Thus, for continuous choice decisions we use the maximum local mass utility function, which captures the idea that there is one best answer and many equally wrong answers, but also avoids the impossible pitfalls of assuming a delta function as the utility structure of the task.

Given the maximum local mass utility function, the optimal observer should choose the mean of the gaussian probability distribution describing her uncertainty; and the sample-based agent should choose the mean of the obtained set of samples (this holds in so far as the utility function and posterior are unidimensional, unimodal, and symmetric; for multi-dimensional problems, see Brainard & Freeman, 1997 for approximation algorithms).

2.6.2 How bad are continuous sample-based decisions?

Now that we know what the optimal and sample-based agents will choose, we can ask: how much worse will the choices made by the sample-based agent be. Figure 2-12 shows the increase in squared error (distance) from the correct answer for the sample-based agent compared to the optimal agent. Since both agents are choosing the mean of the probability distribution, the increased error for the sample-based agent arises from poor estimates of the mean, and this additional error drops off geometrically in the number of samples.

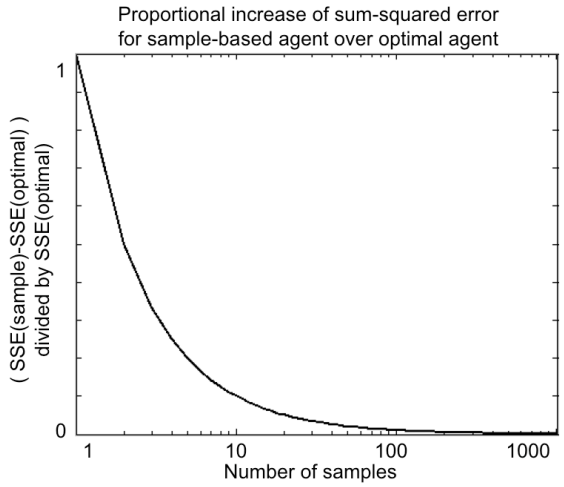

Figure 2-12: Expected increase in error (squared deviation from the optimal answer) for a sample-based agent over the optimal agent about a continuous variable distributed according to a Gaussian probability distribution. With more samples drawn from the Gaussian, the sample-based agent converges to the same decision as the optimal-based agent: the mean of the distribution.

2.6.3 How many samples should the sample-based agent use?

Error rates as quantified by the squared distance from the target are quite meaningless, since these do not take into account the utility function. Thus, we skip directly to an analysis of how many samples will maximize the rate of return for the sample-based agent.

The optimal number of samples for a continuous decision will depend on two factors we had not previously considered. First, the breadth of the distribution predicting the target location, parameterized by its standard deviation, σ_P. Second, the breadth of the utility function, parameterized also by its standard deviation σ_U: how close to the target center does our response have to be to be rewarded. The optimal number of samples turns out to be a function of the ratio between these two standard deviations.

When σ_P is much larger than σ_U, then no matter how many samples we take (to obtain an accurate estimate of the mean of the predictive distribution) our prediction will still be so uncertain that the correct answer is unlikely to be close enough to the mean to be rewarded. Asking how many samples we should take in this case is like asking, "how carefully should we aim when throwing a crumpled piece of paper from the Empire State building into a trash can on the ground?" Obviously, we should not aim very carefully, because no matter how carefully we aim, our success will be left to chance. For exactly the same reasons, in such circumstances we should make decisions based on very few samples, since additional samples will be of no use.

When σ_U is much larger than σ_P, then it also does not make sense to take many samples. When this is the case, if we take one, or infinitely many samples, our guess is

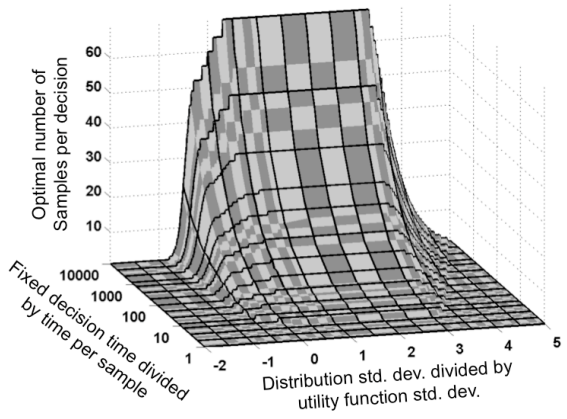

Figure 2-13: Expected increase in error (squared deviation from the optimal answer) for a sample-based agent over the optimal agent about a continuous variable distributed according to a Gaussian probability distribution. With more samples drawn from the Gaussian, the sample-based agent converges to the same decision as the optimal-based agent: the mean of the distribution.

still guaranteed to be near the peak of the utility function, and we will obtain similar rewards. This case is analogous to throwing a piece of crumpled paper into a large trash can situated an inch under your hand: in this case, it makes no sense to spend time aiming because there simply is no way you could miss.

However, in an intermediate range, when the relationship between σ_U and σ_P is just right, then we should obtain many samples to improve performance. This is the scenario when we must throw our paper ball into a trash can from across the room — it is doable, but it is not so easy that we shouldn't aim — it is *just the right level of difficulty*. In this case we would spend the time to take careful aim and delicately arc the toss. Similarly, in this case, we should take many samples when trying to make a decision.

Figure 2-13 shows the optimal number of samples as a function of sample cost and the log of the ratio between σ_U and σ_P. We see exactly the effects described above — at interim ratios, when samples are cheap enough, we should take many of them (within our tested range, as many as 70). However, when decisions are too hard, or too easy, or the sample cost is not low (when it would take at most 10 samples to double our time per decision), we are best off taking just one sample, and making a guess accordingly. Thus, again, when making continuous decisions, it seems that often the best course of action is to make many quick, imperfect decisions to maximize long-run rewards.

2.7 Strategic adjustment of sampling precision

Thus far, we have shown that under some assumptions, in cases when people try to maximize their expected rate of return, making decisions based on very few samples

is actually optimal. However, based on our analysis, we expect that people would use more samples for decision that have higher stakes or are allotted more time — do people make these predicted, optimal adjustments?

A large prior literature on "probability matching" (Herrnstein, 1961; Vulkan, 2000) has studied a very similar phenomenon in a simpler task. In probability matching, subjects predict the outcome of a trial based on the relative frequencies with which that outcome has been observed in the past. Thus, subjects have direct evidence of the probability that lever A or lever B should be pulled, but they do not seem to maximize; instead, they "probability match" and choose levers with a frequency proportional to the probability of reward. On our account, this literal "probability matching" behavior amounts to making decisions based on one sample, while decisions based on more samples would correspond to Luce choice decisions (Luce, 1959) with an exponent greater than 1.

Since probability matching contains a large body of experimental work, we can use this literature for a preliminary evaluation of a key question: do people adjust the number of samples they use as key parameters of the decision-process change? Shanks, Tunney, and McCarthy (2002) concluded that this is the case from a finding indicating that people tend to adopt an ideal maximizing strategy as more training and reward is provided. We can further test the effect of higher stakes on the apparent number of samples used to make a decision in a more graded fashion within the set of experimental findings reviewed by Vulkan (2000). Specifically, we computed the average stakes of the decisions and an estimate of the number of samples subjects used to make those decisions for each of the studies reviewed in Vulkan (2000).

We measure the stakes of decisions as the difference in expected reward (in dollars) between a probability-matching decision and a maximizing decision. These studies vary in the probability of the alternative most likely to be "correct", p, the reward for a correct response, U_+, and the utility for an incorrect response, U_-. The expected *maximizing* reward for these studies is thus $U_* = pU_+ + (1-p)U_-$, and the expected *probability matching* reward is $U_m = (p^2 + (1-p)^2)U_+ + 2p(1-p)U_-$. The quantity we are interested in – what we refer to as the stakes of the decision – is the advantage of maximizing over probability matching, or $U_\delta = U_* - U_m$: for studies where this number is higher, there is more to be gained by taking more samples.

The Luce choice rule describes the relationship between the probabilities of reward associated with various actions and the frequency with which agents choose these alternatives (see Eq. 2.5; Luce, 1959). On the assumption that agents make decisions by sampling, the Luce choice exponent yields a proxy for the number of samples used in a decision. The exact relationship between the Luce choice exponent (L) and the number of samples (k) used to make a binary decision is analytically intractable, and varies as a function of the true probability p. In all cases, however, as k increases, L increases monotonically. We obtain a proxy for the number of samples used to make decisions as the Luce choice exponent of the observed probability matching behavior. If the frequency with which subjects choose the option most likely to contain the reward is p_s, and the probability that the most likely option is rewarded is p_e, then

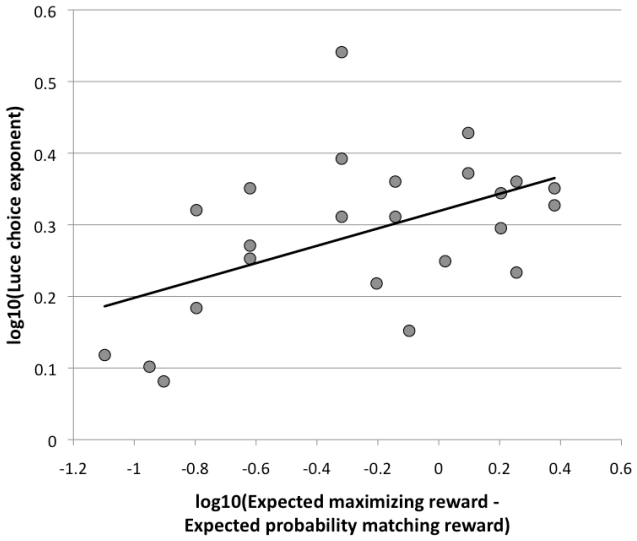

Figure 2-14: Luce choice exponent (a proxy measure for the number of samples used in a decision) as a function of the reward structure (the expected reward from *maximizing* decisions minus the expected reward (in US cents) from *probability matching* decisions. Because both quantities are bounded at 0, we plot their logarithms against each other. Each data point corresponds to one study as surveyed by Vulkan (2000) – despite all of the extraneous variation between studies, a significant correlation is observed: r=0.5, p=0.012.

the Luce choice odds ratio can be described as

$$\left(\frac{p_e}{1-p_e}\right)^L = \frac{p_s}{1-p_s}, \tag{2.5}$$

where L is the Luce choice exponent. Solving for L, we get:

$$L = \log\left(\frac{p_s}{1-p_s}\right) / \log\left(\frac{p_e}{1-p_e}\right). \tag{2.6}$$

With this expression we can measure the Luce choice exponent, L, which is a proxy for the number of samples used in a decision.

Thus we computed the stakes and a proxy for the number of samples used in each of the 24 studies reviewed by Vulkan (2000) that tested probability matching with symmetric, non-zero utility functions. We then measured the correlation between the logarithms of these two quantities (we used logarithms because both quantities are effectively bounded at zero and are not normally distributed otherwise) in Figure 2-14. Our prediction is that when the stakes are higher (that is, when the difference in expected rewards between the maximizing and probability matching response strategies is large) subjects would use more samples for each decision, and thus would show a higher Luce-choice exponent. This is precisely what we find – the stakes and the Luce choice exponent are positively correlated: r=0.5, p=0.012, df=22. Thus, despite all of the other variation across studies, labs, an so on, when stakes are higher, people are closer to maximizing — they seem to use more samples per decision when it matters more.

2.8 Discussion

We began with the observation that, on average, people tend to act consistently with ideal Bayesian inference, integrating information to optimally build models of the world; however, locally, they appear to be systematically suboptimal, acting based on a very limited number of samples. This has been used to argue that people are not exactly Bayesian (Mozer et al., 2008). Instead, we have argued that sample-based approximations are a powerful method for implementing approximate Bayesian inference. Although with few samples, sample-based inferences will deviate from *exact* Bayesian inference, we showed that for choices among 2, 4, 8, 16, and 32 discrete alternatives and for unidimensional continuous choices, a decision based on a very small set of samples is nearly as good as an optimal decision based on a full probability distribution. Moreover, we showed that given reasonable assumptions about the time it takes to produce an exact sample, a policy of making decisions based on very few samples (even just one) is globally optimal, maximizing long-run utility for choices among discrete alternatives as well as choices along continuous variables. Furthermore, our analysis predicts that when the stakes are higher, subjects should use more samples for a decision, and we found evidence of such optimal meta-cognition in a meta-analysis of the probability matching literature.

2.8.1 Related arguments

Other authors have invoked various kinds of sampling as a way to explain human decision making. N. Stewart et al. (2006) suggested that a policy of making decisions through binary preference judgments among alternatives sampled from memory can account for an assortment of human judgment and decision-making errors. Schneider, Oppenheimer, and Detre (2007) suggest that votes from sampled orientations in multi-dimensional preference space can account for violations of coherent normative utility judgments. A promising direction for future research would be to relate models like these, based on samples drawn from memory or over preferences, to models like those we have described in our paper, in which samples are drawn from probability distributions reflecting ideal inferences about the world.

2.8.2 Internal vs. External information gathering

The literature on drift-diffusion modeling, or Weiner/Ratcliff decision processes (Ratcliff, 1978; Gold & Shadlen, 2000), has explored similar questions. In these experiments and models, participants are exposed (or are assumed to be exposed) to noisy input from which they continuously gather information about the world. At some point, this evidence exceeds the threshold required for one of two responses, and subjects make this response. These cases may be construed as obtaining "samples" from the external world, and are thus in some ways analogous to our analyses of sampling from internal beliefs. It has been suggested that people adopt nearly optimal (with respect to maximizing the rate of return) decision criteria in these tasks (Bogacz, 2007; Gold & Shadlen, 2002; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). Whether or not people are exactly optimal, one thing is clear: people are inclined to gather little evidence, and make decisions quickly, rather than spend a lot of time "sampling" evidence (Hertwig & Pleskac, 2010). Thus, it seems that whether decisions are being informed by external information gathering, or internal deliberation, the tradeoff between making quick, less-informed decisions, and slow, more-informed decisions is similar. In both cases, people seem to choose a globally optimal policy of using few samples to make quick decisions.

2.8.3 What is a sample?

What is a sample as we consider it, and what does it take to produce such a sample? There are two important points to be made here. First, to pre-empt an objection from experts in sampling who know that one typically needs many samples for inference, we want to clarify that a sample, as we consider it, is an *exact, independent* sample – that is, a sample from the actual posterior probability distribution. Various approximate sampling schemes, such as *Gibbs* sampling (Geman & Geman, 1984), Markov Chain Monte Carlo (Robert & Casella, 2004), importance sampling (Srinivasan, 2002), or particle filters (Doucet et al., 2001) produce correlated samples. For instance, to produce an exact sample from a Markov Chain Monte Carlo algorithm one must run the algorithm for a fairly long "burn-in" period – in effect, what we consider one

sample, would require many MCMC iteration – and even after the burn-in period, subsequent samples are still correlated. All of these approximate sampling methods are associated with schemes for estimating the effective sample size. In MCMC, this amounts to the auto-correlation of the sampling chain; in importance sampling and particle filters, this is computed from the variance of the importance weights. We expect that these schemes for estimating the effective sample size will yield numbers that can link these more sophisticated sampling algorithms to the analyses we present in this paper.

Second, where does deductive, procedural thought come into play if we cast thinking in terms of sampling? Here we want to clarify that the process for producing one exact sample from a complex, structured model of the world will often require much deductive, procedural calculation. Thus, although we cast the output of this processing as a "sample", the processing itself will often contain many deterministic calculations. For these reasons – that a sample is actually the output of potentially very complicated calculations – we believe that the process of producing one such exact sample is likely to be a rather slow process.

2.8.4 Sample cost

How much might a sample "cost"? In our analyses the relevant measure of sample cost in multiple-trial experiments is the ratio between the time it takes to make an action and go on to the next trial and the time required to draw a sample to inform a decision about that action – a measure of how much using a sample will slow down the rate of decision-making. Ratios near 10 seem quite reasonable: most experimental trials last a few seconds, and it can arguably cost a few hundred milliseconds to consider a hypothesis. This is speculation, however, it seems to us that in most experimental tasks, the benefits gained from a better decision are relatively small compared to the costs of spending a very long time thinking. So, if thinking amounts to sampling possible alternatives before making a decision, it should not be surprising that people regularly seem to use so few samples. Though we have focussed on the time-cost of samples, similar results will hold if samples are costly in energy, or other resources.

2.8.5 Assumption of a uniform prior over p

Our analyses, particularly in the alternate-forced choice domains, have assumed that the sample-based agent assumes an uninformative prior about the problem – that is, the sample-based agent assumes a uniform prior over p in the Bernoulli case. On this assumption, our calculations of the optimal number of samples seem most robust and general; however, the optimal number of samples will vary if the structure of the problem confers a more informative prior. If we assume that we are in a nearly-deterministic setting where p tends to extreme values (0 or 1), then the optimal number of samples will change: these cases guarantee that the first sample will be informative and the second sample will be redundant. On the other hand, if we assume we are dealing with a very random, unconstrained problem, where p tends to be around 0.5, then we know that all samples will be uninformative. If we think

that p tends to be around 0.7 – a regime where more samples pay off, then we would assume that we should use more samples. As assumptions about p will vary, the shape of the optimal number of samples as a function of sample cost will vary; however, only under very constrained conditions will the optimal number of samples be much higher than our analyses have described.

2.8.6 Black Swans and variable utility functions?

What happens with variable utility functions? In our analysis we have assumed that utility functions are constant in nAFC decisions – one reward is assigned for a "correct" answer, and another for an "incorrect" answer. This assumption holds for the bulk of psychological experiments, and even for most methods of evaluating machine learning algorithms; however, it does not apply universally in the real world, where some outcomes are better than other positive outcomes. Although little about our results will change when such variation in utilities is small, it poses an interesting problem when this variation is large.

Take, for instance, the game of "Russian roulette", in which one bullet is placed within a six-shot revolver, and the drum is randomly spun; the player then aims the revolver at their head, and pulls the trigger. In this game, there is a 5 in 6 chance that the current chamber does not contain the bullet, and pulling the trigger will cause no harm, and will confer a slight reward (the game is said to have been played by 19th-century Russian military officers to demonstrate their bravado to others). However, there is a 1 in 6 chance that the chamber does contain the bullet, in which case the loss is catastrophic (death). In these cases, and others, the sample based agent that relies on few samples might not consider the state of the world in which the bullet is in the current chamber. Thus, the agent will not consider the relatively low probability of an extreme outcome. This is referred to as the "Black Swan" problem (Taleb, 2008): Ignoring very important but low probability events leads to substantial biases and irrationalities, which Taleb (2008) argues exist in finance.

One possible method that sample-based agents may adopt to avoid the black swan problem is increasing the sampling rate of high-stakes scenarios. For instance, instead of sampling from just the posterior probability distribution over possible world states, one might weight the samples by the variance of possible outcomes in that state. Using this modified sampling strategy, world-states in which decisions are particularly high stakes will be over-represented relative to their probability, but will allow the agent to compute the expected utility of a particular action in a more useful manner. Such a sampling scheme predicts some forms of availability effects (Tversky & Kahneman, 1974) — mental over-representation of the possibility of events with extreme outcomes. It will be an interesting direction for future research to assess how availability may be used to overcome the black swan problem for sample-based agents and whether this sampling strategy underlies human decision-making biases.

2.8.7 Limitations

We should emphasize that we are not arguing that all human actions and decisions are based on very few samples. The evidence for sampling-based decisions arises in high-level cognition when people make a decision or a choice based on what they think is likely to be true (Which example is in the concept? How long will this event last? How many airports are there in the US?). In other situations people appear to integrate over the posterior, or to take many more samples, such as when people make graded inductive judgments (How similar is A to B? How likely is it that X has property P given that Y does? How likely do you think that F causes G?). Moreover, in low-level sensory and motor tasks, decisions often seem to be much closer to ideal Bayesian performance, rather than decisions based on few samples, as seen in cognition (Trommershauser, Maloney, & Landy, 2003 although see Battaglia & Schrater, 2007). It is interesting to consider why there may be a difference between these sorts of decisions and tasks.

2.8.8 Conclusion

Under reasonable discrete and continuous choice scenarios, people are best advised to make decisions based on few samples. This captures a very sensible intuition: when we are deciding whether to turn left or right at an intersection, we should not enumerate every possible map of the world. We do not need "moral certainty" about the probability that left or right will lead to the fastest route to our destination—we just need to make a decision. We must implicitly weigh the benefits of improving our decision by thinking for a longer period of time against the cost of spending more time and effort deliberating. Intuition suggests that we do this in the real world: we think harder before deciding whether to go north or south on an interstate (where a wrong decision can lead to a detour of many miles), than when we are looking for a house (where the wrong decision will have minimal cost). Indeed, empirical evidence confirms this: when the stakes are high, people start maximizing instead of "probability matching" (Shanks et al., 2002), and we show that they do so in a graded fashion as stakes increase. Nonetheless, it seems that in simple circumstances, deliberating is rarely the prudent course of action—for the most part, making quick, locally suboptimal, decisions is the globally optimal policy: one and done.

Chapter 3

Attention as inference: Selection is probabilistic; Responses are all-or-none samples

3.1 Thesis framing

So far I have argued from a purely theoretical perspective that people can approximate Bayesian inference and decision-making by using sample-based approximations with only a few samples. In this chapter, I will use common visual selective attention tasks – which allow the precise stimulus-response control – to assess whether individual guesses from individual subjects on individual trials have the statistical properties of independent samples.

This chapter was published as: (Vul et al., 2009)

Abstract

Theories of probabilistic cognition postulate that internal representations are made up of multiple simultaneously-held hypotheses, each with its own probability of being correct (henceforth, "probability distributions"). However subjects make discrete responses and report the phenomenal contents of their mind to be all-or-none states rather than graded probabilities. How can these two positions be reconciled? We recast selective attention tasks such as those used to study crowding, the attentional blink, RSVP, etc. as probabilistic inference problems, and use these tasks to assess how graded, probabilistic representations may produce discrete subjective states. We asked subjects to make multiple guesses per trial, and used second-order statistics to show that: (a) visual selective attention operates in a graded fashion in time and space, selecting multiple targets to varying degrees on any given trial; and (b) responses are generated by a process of sampling from the probabilistic states that result from graded selection. We conclude that while people represent probability distributions, their discrete responses and conscious states are products of a process that samples from these probabilistic representations.

3.2 Introduction

Physical constraints prevent us from producing multiple different actions at once, action is necessarily all-or-none. No matter how unsure we are whether to turn left or right, we can only move in one direction. And no matter how unsure we are of our beliefs, we can only vocalize a single utterance. This all-or-none constraint on human action is so obvious that we often build it into real-world decision procedures (e.g., voting) and we design our experiments around it (e.g., N-alternative forced choice). It is not only our actions, but also our conscious states that seem to be all-or-none: a Necker cube appears to be in one configuration or another, never in both simultaneously. Researchers have attempted to circumvent all-or-none reporting constraints by using Likert scales to tap into graded phenomenal experience. But even when people are asked to report graded degrees of awareness, they use the available scale in an all-or-none fashion, reporting that they are either aware or not aware, rarely "half-aware" (Sergent & Dehaene, 2004).

Such introspections have resulted in many all-or-none accounts of cognitive representation. We consider all-or-none representations to be those that consist entirely of Boolean valued beliefs, i.e., beliefs that are either true or false, but not in-between. In choices from multiple discrete options, one or more options may be deemed true, and the others false. An object either belongs to a category, or it does not; a signal has either passed threshold, or it has not. In choices along continuously valued dimensions (e.g., brightness), all-or-none representations take the form of point-estimates (e.g., 11.3 Trolands). Although the content of a point-estimate is continuous (11.3), its truth value is all-or-none (e.g., "it is true that the brightness of the signal was 11.3 Trolands"). Such all-or-none accounts of mental representation have been postulated for signal detection (point estimates corrupted by noise; e.g., Green & Swets, 1966), memory (memory traces as point estimates; e.g., Kinchla & Smyzer, 1967), concepts and knowledge (as logical rules and Boolean valued propositions; e.g., Bruner et al., 1956).

However, other theoretical perspectives treat mental representations as probability distributions, in which multiple alternative hypotheses are held simultaneously, each with a different graded truth probability. According to one recent framework for modeling cognition, mental tasks can be optimally solved by Bayesian inference (Chater & Oaksford, 2008; Chater, Tenenbaum, & Yuille, 2006). Indeed, a variety of experiments show that human behavior often reflects this optimality, which implies that people are doing something like Bayesian inference (Kersten & Yuille, 2003; Steyvers et al., 2006). Implicit in the claim that people perform Bayesian inference is the idea that human cognitive machinery operates over probability distributions that reflect the uncertainty of the world (Chater et al., 2006; Griffiths & Tenenbaum, 2006). Representations of probability distributions are not all-or-none Boolean values, but rather graded probabilities: every possible decision (left or right), estimate (amount of light present), or state (Necker cube tilted up or down), is assigned a probability that may be any value between 0 and 1. Probabilistic accounts have been proposed for memory (Steyvers et al., 2006), signal detection (Whitely & Sahani, 2008), categorization (Tenenbaum, 1999), and knowledge (Shafto et al., 2008; Vul &

Pashler, 2008).

Although these probabilistic accounts have recently gained much favor in cognitive science for their mathematical elegance and predictive power (Chater & Oaksford, 2008), they conflict with the common intuition that conscious access is all-or-none. How can we have both probabilistic representations, and seemingly all-or-none conscious experience?

We will tackle the conflict between all-or-none subjective experience and probabilistic accounts of representations within the domain of visual selective attention. This domain is an ideal testing ground for several reasons. First, irrespective of debates about cognitive representation broadly construed, the representation underlying visual selective attention has been disputed, with some postulating all-or-none, Boolean representations (Huang & Pashler, 2007; Huang, Treisman, & Pashler, 2007) and others suggesting graded representations (Reeves & Sperling, 1986; Shih & Sperling, 2002). Second, probing the fine line between conscious access and unconscious representation requires a domain that examines that interface: although the link between conscious access and visual attention has long been discussed and debated (Baars, 1997; Koch & Tsychiya, 2007; Lamme, 2003; Posner, 1994), the only clear consensus is that they are closely related. Finally, visual selective attention tasks are appealing because they afford precise manipulations and rigorous psychophysical measurements.

Thus, we will use visual selective attention tasks here to study internal (short-term memory) representations and how subjects use them. First, we will provide a theoretical framework, casting a large class of attentional selection tasks as problems of inference under uncertainty. We will then describe experiments that test whether visual selective attention produces all-or-none representations, or graded representations, akin to the probability distributions implicated in Bayesian inference. Our evidence supports the latter view, and suggests that conscious responses constitute all-or-none samples from these probability distributions.

3.2.1 Visual selective attention

The term "visual attention" encompasses many disparate phenomena sharing the feature that people can selectively distribute resources among the elements of the visual world: e.g., memory (Chun & Potter, 1995; Vul, Nieuwestein, & Kanwisher, 2008)), perceptual fidelity (Carrasco, 2006; Posner, Snyder, & Davidson, 1980), feature integration (Treisman & Schmidt, 1982), and object formation (Kahneman, Treisman, & Gibbs, 1992). Here we consider a class of tasks in which subjects are directed by a cue to select one or more elements for subsequent report (thus allotting memory capacity preferentially to some items over others). In a classic example of such a task, people are presented with a rapid serial visual (RSVP) stream of letters, one of them is cued (e.g., by virtue of being surrounded by an annulus), and the subject must select that letter, remember its identity, and report that letter identity later. Similarly in spatial selective attention tasks, an array of letters may be presented in a ring around fixation, with one of them cued for subsequent report by a line (see Figure 1).

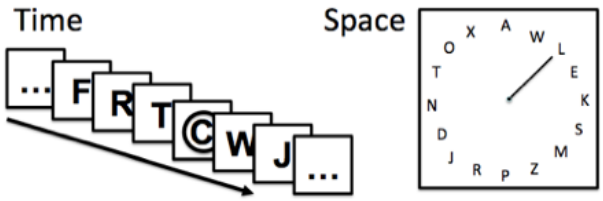

Figure 3-1: Prototypical experimental paradigms used in attentional selection tasks in time and space. One of many items is cued, and the subject must report that item.

Such tasks have been used to study the attentional blink (Chun, 1994, 1997; Chun & Potter, 1995; Raymond, Shapiro, & Arnell, 1992; Nieuwestein & Potter, 2006; Vul, Hanus, & Kanwisher, 2008; Vul, Nieuwestein, & Kanwisher, 2008), crowding (He, Cavanagh, & Intrilligator, 1996; Pelli, Palomares, & Majaj, 2004; Strasburger, 2005), illusory conjunctions (Prinzmetal, Henderson, & Ivry, 1995; Prinzmetal, Ivry, Beck, & Shimizu, 2002), change detection (Landman, Spekreijse, & Lamme, 2003), and short-term memory (e.g., partial report; Averbach & Coriell, 1961). In these experiments, researchers measure which items were reported and infer the properties of attentional selection (e.g., when it fails, and what its limits are). Rather than investigating the limits of attention in such tasks, here we are primarily concerned with the output of the selection process: the representation in short-term memory that attentional selection creates on any given trial of such an experiment.

Two main classes of theories address the issue of the representation that attention produces when selecting a particular object or region for storage in memory and subsequent report. According to one theory, items are selected through an attentional gate that defines a weighting function in space and time (Shih & Sperling, 2002). Therefore, on this account, the short-term memory representation resulting from selection is a weighted list with items closer to the cue receiving a higher weight, and those further from the cue receiving a lower weight. A contrasting recent theory postulates that items are selected by a Boolean Map that defines some spatial regions as wholly selected, and others as not selected, but does not include graded weights, or half-selected regions (Huang & Pashler, 2007). Therefore, on this account, the representation of possible items in short term memory should be Boolean an item will be either within the selected region, and remembered, or outside the selected region, and forgotten as a non-target.

3.2.2 Selective attention as inference under uncertainty

Theories of attention are usually cast at an algorithmic level, but it is also useful to consider Marrs (1982) computational theory level of explanation by asking what are the problems being solved in these tasks. Bayesian inference provides a useful framework, enabling us to relate attentional selection to probabilistic cognition. Several groups have recently posed Bayesian accounts for mechanisms of attentional enhancement (Yu & Dayan, 2005), deployment of attentional enhancement or eye movements

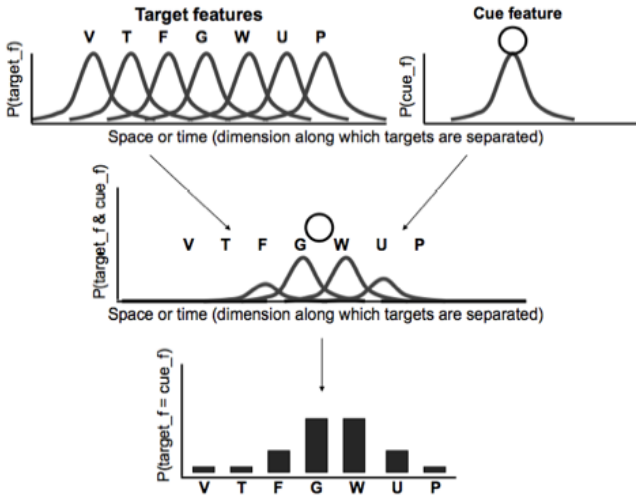

Figure 3-2: Selective attention as inference. Typical experimental paradigms make the task difficult by capitalizing on uncertainty in the spatial or temporal position of the targets and cues. The task thus amounts to inferring which target was likely to have co-occurred with the cue. A Bayesian solution to this task would result in a probability distribution over possible targets.

(Itti & Baldi, 2009; Najemnik & Geisler, 2005), or the integration of top-down influences with bottom-up saliency-maps (Mozer, Shettel, & Vecera, 2005). Here we apply the probabilistic approach to the "attentional selection" tasks we have discussed, and cast these tasks in terms of inference under uncertainty.

What problem is being solved by visual selective attention in these tasks? Specifically, we want to know what the output of the attention mechanism ought to be given the nature of the problem. In a typical experiment on attentional selection, that problem entails reporting one feature or object (a "target"; e.g., the letter identity, "A") that is distinguished from distracter items by some "cue" (e.g., an annulus) a stimulus that identifies the spatial or temporal location of the target (Figure 1). The spatiotemporal location is simply one or more dimensions along which different items are arrayed. Thus, attentional selection tasks amount to assessing which of the potential targets spatially or temporally co-occurs with the cue and then allocating short-term memory based on the solution. To make the task challenging such that informative patterns of failure may be observed, the experimenter controls the discriminability of possible targets in time or space by taxing the system in different ways (e.g., close spatial or temporal packing of targets, brief display durations, etc). These conditions introduce spatial and temporal uncertainty about the locations of each possible target, as well as the cue.

The subjects task, then, is to determine which target coincided with the cue, given some uncertainty about the spatiotemporal locations of both the target and the cue. This task may therefore be considered inference under uncertainty, which is optimally solved by Bayesian inference. Given particular levels of uncertainty, the

Bayesian solution to this problem entails, for each item, and point in time, multiplying the probability that the letter occurred at that point in time, by the probability that the cue occurred at that point in time, and then integrating over time to obtain the probability that this letter coincided with the cue. The solution to this co-occurrence detection problem is a probability distribution over items, describing the likelihood that each item coincided with the cue. If this description is correct, and attentional selection is indeed solving the inference problem just described, it should produce probability distributions over items likely to be the target (see Figure 2). We will test whether people represent such a probability distributions over items in short-term memory.

3.2.3 Within-trial gradation, across-trial noise, and representation

The typical experimental design in cognitive psychology precludes researchers from determining whether internal representations were all-or-none or graded on any one trial. The problem is caused by averaging across trials and subjects (e.g., Estes, 1956). Consider the task of reporting a cued letter from an RSVP sequence of letters. Subjects will not report the target correctly on all trials, but will sometimes instead report the letter before or after the target, or occasionally another letter even farther away in the RSVP sequence (Botella, Garcia, & Barriopedro, 1992; Kikuchi, 1996) . A histogram of such reports across trials will show a graded variation in the tendency to report items from each serial position (see Figure 3, bottom row), as expected given the uncertainty inherent in the task.

It is tempting to interpret this graded variation as indicating that selection itself is graded (Botella & Eriksen, 1992; Reeves & Sperling, 1986; Weichselgartner & Sperling, 1987). However, this conclusion does not follow, because variation in the items reported might reflect not gradations in the degree to which each item is selected on any given trial, but rather variation across trials in which items are selected. That is, the graded across-trial averages are consistent with the possibility that on each trial subjects select items in an all-or-none-fashion, but which items are selected varies across trials due to variability, or noise, in the deployment of attention. This distinction is analogous to the classic dichotomy in signal detection theory: is the variability in whether a stimulus is reported as visible due to noise that varies across trials (Green & Swets, 1966; Nieuwestein, Chun, & Lubbe, 2005), or uncertainty that is represented on every trial (Vul & Pashler, 2008; Whitely & Sahani, 2008). Thus, across-trial histograms are not indicative of the properties of selection on any given trial.

Logically, the observed distribution of reports across trials is the combination of the across-trial variance and the within-trial gradation of selective attention. Figure 3 shows a few of these possibilities if the within-trial spread and across-trial gradation are both Gaussian. Within-trial gradation refers to the properties of selection on any one trial: that is, the representation in short-term memory resulting from selection. Across-trial variance, on the other hand, corresponds to the properties of this rep-

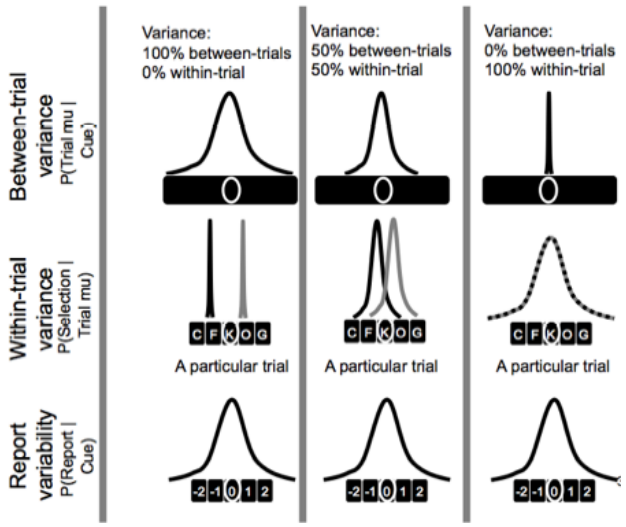

Figure 3-3: The final distribution of reports across trials is a combination of across-trial variation in where/when attention is deployed, and the within-trial gradation in the extent to which an item is selected on a given trial.

resentation that change across trials. That is, given the within-trial distribution of selection on any given trial, how does it vary from one trial to the next (due to noise, or other factors)?

There are an infinite number of plausible combinations of within-trial gradation and across-trial variability in selection that could produce the same final pattern of results. The experiments presented in this paper rule out many of these possibilities, but a few alternatives will remain. Before describing our experiments, it is worth laying out a few qualitatively different cases.

Single-item selection: The simplest alternative is that subjects select and store in short-term memory only one letter per trial. For our purposes, this representation is essentially all-or-none, since only one letter is stored. On this account, the gradation in the frequency that a particular item is reported across all trials corresponds entirely to across-trial variability, which may be characterized as translation of the single-item selection window.

Contiguous all-or-none selection: Subjects may select multiple spatiotemporally contiguous items in an all-or-none fashion (an item is either selected, and stored in memory, or not). On this account, there must be across trial variance to produce a graded distribution of final reports in the across-trial average. Furthermore, because the selected items are contiguous, this variability can only take the form of spatiotemporal translation of the selection window.

Contiguous graded selection: The alternative we advocate is that multiple contiguous items are selected, but the degree of selection varies across items on a given trial. This defines a weighting function over items, which may be described as a graded "attentional gate" (Shih & Sperling, 2002), or a probability distribution. On this account, there may also be a spatiotemporal translation from trial to trial (as shown

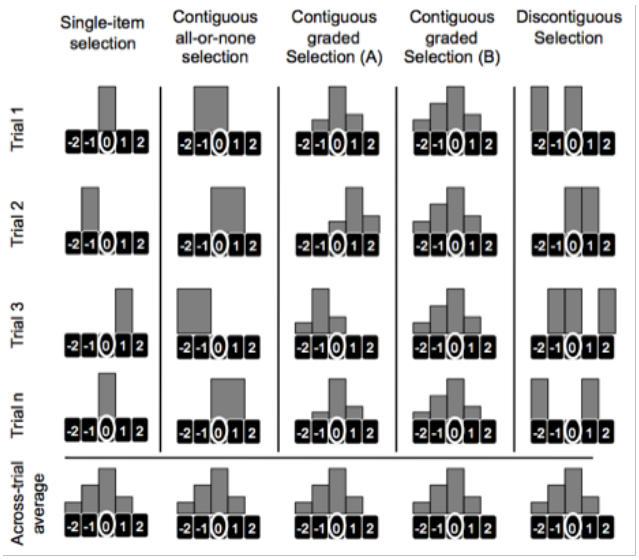

Figure 3-4: Several possible modes in which attentional selection may be operating to produce across-trial variation (described in more detail in text).

in Figure 3), but that is not necessarily the case, since the gradation of the selection window in this case may perfectly match the gradation seen in the final distribution of reports, as would be the case if there were no variability across trials.

Complex selection: One final possibility is that the items selected on a given trial need not be in a fixed spatiotemporal relationship (e.g., contiguous), as would be the case if several independent single-item selection episodes occurred on each trial. If this is the case, then across-trial variability is not constrained to be a mere translation of the attention window: it may take any form (e.g., on trial A, items -2, 0, and 1 are selected; on trial B, items -1 and 0 are selected). This account could describe any set of data, because both (a) all other accounts are its subsets, and (b) it allows for as many degrees of freedom as there are data points. A similarly complicated account is that subjects may select a variable number of contiguous items, thus the final response distribution will be a mixture of uniform distributions defined over different intervals of the presented items this account also is sufficiently unconstrained that it could account for any pattern of data (by postulating specific proportions for different components of the mixture). As we will describe later, our data suggest a more parsimonious account. For these reasons, we will not consider these alternatives again until the discussion.

3.2.4 Within-trial representations, attention, and probability

The alternatives above propose different amounts of "within-trial gradation" and "across-trial variability" of representations. Within-trial gradation (rather than across-trial variability) of representations has implications for selective attention, as well as

probabilistic representation more broadly. First, within-trial gradation can determine whether selective attention operates in a graded or discrete fashion. Evidence for any amount of within-trial gradation of selection would conflict with recent theories of spatial selection that suggest that selection operates as a Boolean map, selecting regions of space in an all-or-none fashion (Huang & Pashler, 2007). On the other hand, evidence for any amount of across-trial variability in selection would call into question previous research using the distribution of reports across trials to infer the properties of selection on any one trial (Shih & Sperling, 2002).

Second, within-trial variability is also a measure of how people represent uncertainty on any given trial. A substantial amount of within-trial variability implies that subjects represent the uncertainty inherent in a particular task on every trial. This finding would suggest that internal representations may indeed be probability distributions. However, if we find only across-trial variability in reports, our results would suggest that many previous results showing that responses follow probability distributions appropriate to the inference in question, may be an artifact of averaging across people or trials the probability distributions exist across individuals or time, but not within one individual at a specific point in time (Mozer et al., 2008).

Most importantly, if we find that attention operates in a graded fashion, the results will have ramifications beyond the realm of visual selective attention to the nature of perceptual awareness. Introspection, as well as some data, suggest that awareness is discrete: We are either aware or something, or we are not. Sergent and Dehaene (2004) tested this intuition by asking subjects to provide ratings of their degree of awareness of the target item in an "attentional blink" (Raymond et al., 1992) paradigm. Subjects reported bimodal degrees of visibility: sometimes the target was rated as completely visible, sometimes completely invisible, but participants rarely provided intermediate ratings. These results suggest that conscious access may be a discrete phenomenon. A similar conclusion was reached from studies of the wagon wheel illusion under continuous light. In movies, a rotating wagon-wheel can appear to move backwards due to aliasing arising from the discrete sampling frequency of the movie frames. Because the wagon-wheel illusion can be seen under continuous light, some have argued that perception is discrete: the wagon-wheel moves backwards due to aliasing arising from discrete sampling of percepts from the environment (Van Rullen & Koch, 2003). Given these findings, if the present studies find that selective attention is continuous, in that it produces graded representations, we must reconcile this fact with the apparent all-or-none nature of conscious awareness.

In the experiments reported here, we measure the across-trial variance and within-trial spread of selection by asking subjects to make multiple responses on a given trial: subjects first report their best estimate of the item that was cued, and then make additional guesses about which item was cued. This method has been used previously in research on signal detection theory (Swets, Tanner, & Birdsall, 1961), and more recently to study representations of knowledge (Vul & Pashler, 2008). As in this previous literature, we consider the relationship between errors on the first response, and the second response. In our case, we consider the position of items reported in a selective attention task, and evaluate whether two items reported on one trial are independent (as predicted if they are samples from a probability distribution), or

whether they share some variance (as predicted from across-trial variability). For example, if subjects incorrectly report an item appearing earlier in the RSVP list as the target, will a second guess from same trial likely be another item that appeared early in the list? If so, then there is some common error for the trial shared across guesses, indicating that there is some across-trial variability in which items are selected (thus giving rise to a graded final distribution of reports). If the temporal positions of the intrusions reported in the two guesses are not correlated, then there is no common, shared, error for a given trial, and the final distribution of reports is driven entirely by within-trial variability.

For single item selection, we don't expect to find information in both guesses (even if the subject postpones reporting the selected item until the second guess, there will be no systematic relationship between the items reported on guess 1 and 2). For contiguous all-or-none selection to produce a graded final distribution of reports, variability must exist in the position of the selection window across trials. This translation would necessarily induce a correlation in the errors of two responses, and thus the contiguous all-or-none selection account mandates a correlation. Only the contiguous graded selection account can produce a graded final distribution of reports without any across-trial variation (and thus correlation of errors).

Thus we test for within- and across- trial variability of temporal selection in Expeirment 1, and of spatial selection in Experiment 2. In both cases we find that the there is no correlation in the temporal or spatial position of intrusions from multiple responses on one trial. This finding indicates that there is no across-trial variability, and therefore, the average distribution of final reports reflects the gradation of selection on any given trial. Thus, selection is continuous and graded, while responses act as samples from the graded representation. Our data indicate that attention selects a number of items to varying degrees on any given trial, creating a probability distribution over likely targets, and subjects make responses and subjective judgments by sampling items from the selected distribution while having no conscious access to the distribution itself.

3.3 Experiment 1

First, we test whether selective attention is graded: are multiple items selected to varying degrees on a given trial, and does this within-trial spread of selection underlie the commonly observed final distribution of reports? Commonly adopted experiments with single-probe trials do not provide enough information to dissociate across-trial variance and within-trial gradation. To assess the spread of the items selected by attention on a given trial, we asked subjects to make four guesses about the identity of the cued target. By analyzing the distributions of subsequent guesses conditioned on the first guess, we can estimate the spread of selection within a given trial.

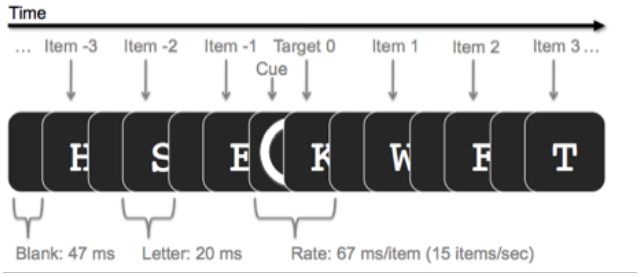

Figure 3-5: Experiment 1 design: subjects were asked to report one item cued in a rapid serial visual presentation, but were asked to make four guesses about the identity of that item. The decision to display the cue before, rather than during, the target was irrelevant to our main analysis of the relationship between multiple guesses on a single trial, and was done to match performance with some unrelated previous work (Vul, Hanus, & Kanwisher, 2008).

3.3.1 Method

Participants. Nine subjects from the Massachusetts Institute of Technology subject pool were recruited to participate. Subjects were between 18 and 35 years of age and were paid $10 for participation.

Materials. On each trial, subjects saw an RSVP stream composed of one instance of each of the 26 English letters in a random order. Each letter was presented for 20 msec, and was followed by a 47 msec blank (3 and 7 frames at a 150 Hz refresh rate, respectively), resulting in an RSVP rate of 15 items/sec. Letters were white on a black background, capitalized, in size 48 Courier font. With our resolution (1024x768), monitor (Viewsonic G90f), and viewing distance (roughly 50 cm), letters subtended roughly 2.5 degrees of visual angle.

On each trial, one cue appeared in the RSVP stream to indicate which of the letters was the designated target. The cue was a white annulus with an inner diameter of 2.8 degrees and an outer diameter of 3.2 degrees; thus the cue appeared as a ring around the RSVP letter sequence. When a cue appeared, it was shown in the 47 msec blank interval between two letters (see Figure 5).

Onset of the cue was randomly counterbalanced to appear either before the 6th, 8th, 10th, 12th, 14th, 16th or 18th letter of the sequence. Subjects were asked to report whatever letter appeared immediately after, or at the same time as, the cue. The experiment was programmed in PsychToolbox (Brainard, 1997) on Matlab 7 on a Windows XP computer.

3.3.2 Procedure

Each participant began the experiment with two practice trials; the results of these trials were discarded. Following the practice trials, participants completed 3 blocks of 70 trials each. Each block contained 10 instances of each of the seven possible cue onset positions, in a random order for each block.

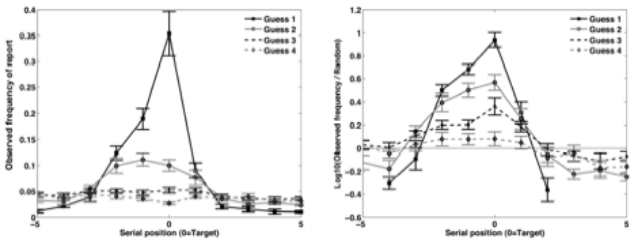

Figure 3-6: Experiment 1 results: (a) The frequency that a given serial position is reported for each of the four guesses subjects make on each trial. (b) The log ratio of the empirical frequency of report, compared to the frequency of chance report: this quantity effectively corrects for the decreasing chance rate of reporting particular items (if they are reported on prior guesses) and corresponds to how much above chance a particular serial position was reported on each guess. Error bars show 1 s.e.m. across subjects.

At the end of each trial subjects were asked to make four guesses about which letter they thought was cued by the annulus. Subjects reported the letters by pressing the corresponding keys on the keyboard. Duplicate letter reports were not accepted, thus each guess was a unique letter.

Subjects were told that they would get 1 point if they reported the letter correctly on the first guess, 0.5 points on the second guess, 0.25 points on the third guess, and 0.125 points on the fourth guess. Feedback and scoring on each trial reflected this instruction. To motivate subjects to perform well on this task, in addition to the flat rate of $10 for participation, subjects were offered bonus cash awards for performance: $0.01 for each point scored (on average subjects scored 160 points in a given session: $1.60 bonus).

3.3.3 Results

Because there were no repeated letters on any trial, we could identify the exact serial position of the reported letters. From this information, we computed the distribution of guessed letters around the presented cue.

Figure 6a shows the empirical frequency with which a letter from each serial position was reported as a function of distance from the cue. That is, an x value of 0 corresponds to the cued letter (target); an x value of -1 is the letter that preceded the target; and an x value of 1 is the letter than followed the target. This is shown for each of the four guesses. The distribution of first guessed serial positions shows a pre-cue intrusion pattern, that is, items preceding the cue (negative serial positions) are reported more often than items after the cued letter (positive serial positions). Effects such as this have been reported before under certain conditions (Botella et al., 1992; Kikuchi, 1996); presumably in our data, these effects are increased because the cue actually appears between the preceding distracter and the target.

Serial positions that are reported above chance may be identified in Figure 6b

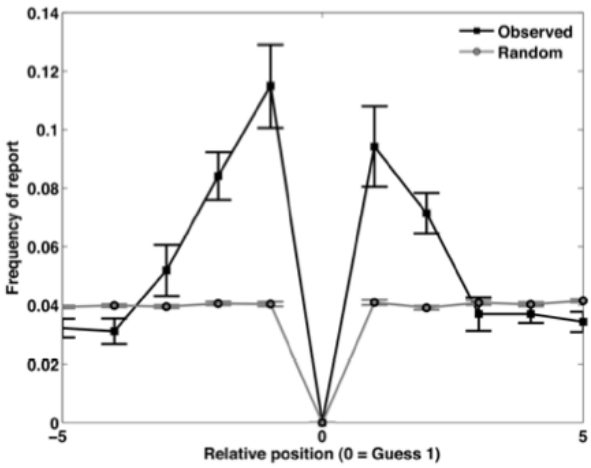

Figure 3-7: Experiment 1 results: The serial position of the item reported on guess 2 relative to the serial position of the item reported on guess 1. Subjects are likely to report two temporally adjacent letters on guesses 1 and 2, indicating that multiple letters are selected on any given trial. Error bars show 1 s.e.m. across subjects.

as those points with log likelihood ratios (log of the empirical frequency divided by chance frequency) above 0 (significance may be ascertained by the error bars, which correspond to 1 standard error of the mean, across subjects). These log likelihood ratios for guesses 2-4 suggest that guesses 2-4 have roughly the same distribution of reports as the first guesses, given that the peak (position 0, target) could not be guessed twice. However, this distribution also has an ever-increasing admixture of random, chance reports. Since guess 3 and 4 are at, or close to, chance, all of our subsequent analyses will look at just guesses 1 and 2.

The fact that guess 2 is above chance would seem to rule out the possibility of single item selection, since to have a reliable second guess subjects must have selected more than one letter. This conclusion may also appear to follow from the observation that subjects produce a similar distribution on guess 2 as guess 1. However, these facts do not indicate how much of the variance seen in the distribution of reports on guess 1 (what is normally measured in such tasks) is attributable to across-trial variance and within-trial gradation. This pattern of results may also arise if, on any given trial, subjects select one and only one letter, but on some trials subjects pressed the wrong key on guess 1, and responded with the actual selected letter on guess 2 (or 3, or 4), thus raising performance on those subsequent guesses above chance. To determine whether this was the case for the second guess, we can look at the distribution of second guess reports relative to the serial position of the first guess report. If subjects only select one letter per trial, and either report it on the first or second guess, there should be no reliable relationship between the serial position of the first guess and the serial position of the second guess.

Figure 7 shows the frequency of guess 2 reports as a function of serial position distance from the letter reported on guess 1. These data show that the second guess

is likely to come from one of the four serial positions nearest to the first guess (these serial positions are reported above chance: all four t values ¿3.3, df = 8, ps¡.005). This indicates that subjects must be selecting at least two letters in proximal serial positions on any given trial. This pattern of results cannot arise from the single item selection account, in which one and only one letter is selected on a single trial. Thus, we can say that multiple items are selected on each trial.

We must ascertain how much of the variance in reports arises from across-trial variability to assess whether the items selected on a given trial are selected in an all-or-none or a graded fashion (the contiguous all-or-none and contiguous graded selection accounts). A graded tendency to report particular serial positions across trials must arise from across-trial variability if selection takes the form of an all-or-none contiguous block on any given trial. However, if selection on a given trial may be graded, then there need not be across-trial variability to produce a graded across-trial report frequency. Thus, the contiguous all-or-none account predicts a substantial amount of across-trial variability, as this is the only way that a graded distribution of errors may arise in the across-trial average.

To measure across-trial variability, we exploited the idea that across-trial variance in the form of temporal translation of the selected region should affect Guess 2 reports and Guess 1 reports similarly, such that Guess 2 reports should depend on the serial position of Guess 1 reports. If there is zero across-trial variance, all guesses are sampled from the same distribution, which corresponds to the degree to which each letter is selected on every trial. Therefore, regardless of the absolute serial position of guess 1, the distribution of absolute serial positions of guess 2 should be unchanged. However, if there is substantial across-trial variance, then the guesses will be sampled from different distributions on different trials. Thus, on trials when Guess 1 was reported as (e.g.) the item two letters before the cue (-2), the distribution of reported Guess 2 serial positions should shift towards -2 (as it is sampled from the same, uncentered distribution as Guess 1). Figure 8 provides an illustration of this conditional-response distribution logic. Thus, we can estimate the across-trial variance by testing whether the distribution of Guess 2 reports is independent of Guess 1 reports.

Figure 9 displays this conditional-report distribution analysis: the distribution of guess 2 reports conditioned on the serial position of guess 1 reports. These conditional distributions are not substantially different from one another: they all appear to be sampled from the same distribution that we see in average Guess 1 reports. A crude way to assess whether guesses 1 and 2 are dependent is to compare the average serial position reported for guess 2 (within the range of -1 to 1) on trials where guess 1 came from serial position -2 to trials where guess 1 came from serial position 2. This comparison shows no significant difference (t(8)¡1), and the 95% confidence interval of the difference straddles 0 (-0.74 to 0.36).

Another test of independence is to evaluate the correlation between guess 1 serial position and guess 2 serial position. To make this test more conservative we consider only trials on which guess 1 and guess 2 came from serial positions -3 through 3, thus we discard most noise trials. Moreover, we discard trials in which subjects report the same absolute-value serial position for guess 1 and guess 2 (e.g., -1 and 1); thus we get rid of the bias that would otherwise exist in this analysis because subjects

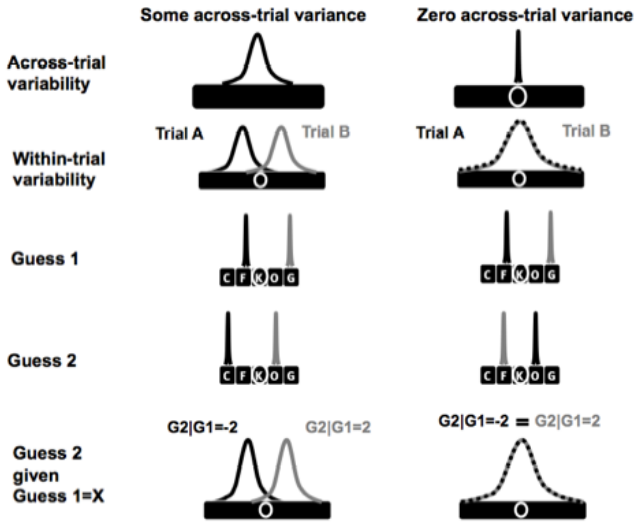

Figure 3-8: The logic behind the conditional analysis in Experiment 1. If there is a substantial amount of between trial variance (column 1), then, on some trials, earlier RSVP positions will be selected (e.g., Trial A, green), and on other trials later positions will be (e.g., Trial B, red). Thus, on this account, guesses 1 and 2 will be dependent, in that a guess from an early serial position of guess 1 would predict more early-stream reports for Guess 2. If there is zero across-trial variability (column 2), then the selected distribution will be identical on every trial, and guesses 1 and 2 will be independent. (See text for further details)

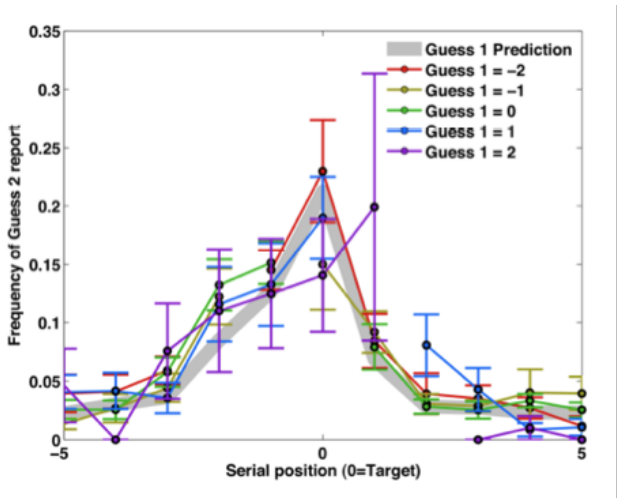

Figure 3-9: Experiment 1 results: Guess 2 reports conditioned on Guess 1. The distribution of guess 2 reports does not change as a function of guess 1 report. This indicates that there is very little (if any) across-trial variance Guess 1 and Guess 2 are independent. Moreover, the distribution of guess 2 reports follows the distribution predicted by guess 1 reports (thick gray line), indicating that guess 1 and guess 2 are identically distributed.

cannot report the same letter twice. This leaves us with an average of 82 trials per subject. This analysis reveals no correlation between response 1 and response 2: an average correlation of -0.06, with 95% confidence intervals across subjects between -0.15 and 0.02 (thus, if anything, there is a negative correlation). Thus, this analysis also shows that guess 1 and guess 2 are independent, with respect to their average serial position, as predicted if there were no (or very little) across-trial variability in the temporal position of the selection window.

Our claim that the conditional guess 2 distributions are unchanged regardless of the serial position that guess 1 came from can be more conservatively tested by asking whether the frequency of reports of any serial position differs between any of the 5 guess-1 conditions. To test this, we computed 30 pairwise comparisons. For instance, one such comparison: probability of reporting serial position 2 on Guess 2 after Guess 1 was serial position 1, compared to the probability of reporting serial position 2 for Guess 2 when Guess 1 was serial position 0. We did such comparisons for every combination of the five Guess 1 report conditions, for Guess 2 reports in every serial position between -2 and 2 (where reports were above chance note that this is more conservative than comparing all of the serial positions, many of which are at chance for all conditions). Of those 30 comparisons, only 2 had p¡0.05, as would be expected by chance. Even if one adopts a lenient correction for multiple-comparisons (Dunn-Sidak), none of the 30 comparisons are significant. Thus, we conclude that the distribution of letters reported in the second guess is independent of the serial position of the first guess. This would not be the case if there was any substantial across-trial variance resulting in different distributions from which reports are sampled trial to

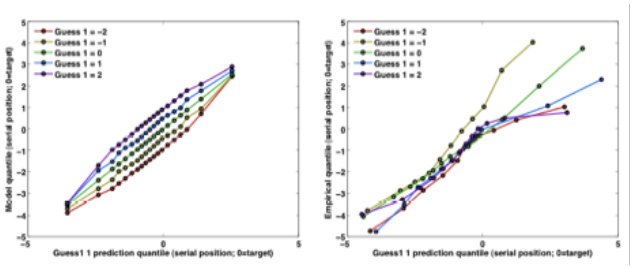

Figure 3-10: Quantile-quantile (QQ) plots that can be used to analyze the similarity between two probability distributions. (a) What the QQ plots would look like if guess 2 shifted in the direction of guess 1 errors. (b) The observed QQ plots, deviations from the diagonal are seen only in the extreme values.

trial. Thus, we conclude that guess 1 and guess 2 errors are independent (Vul & Pashler, 2008).

Finally, we compared these conditional distributions of reports to the distribution we would expect if guess 2 were another independent sample from the same distribution from which guess 1 was drawn. We performed this simulation correcting for the increased rate of random guessing on guess 2 as well as the fact that the same letter could not be reported for guesses 1 and 2. Along with the conditional distributions of report, Figure 9 also shows this guess-1-model prediction (thick gray line). Deviations from the guess-1 predictions are well within the errors of our measurement (R2=0.70, p¡.00001). This further bolsters our claim that all guesses are samples from the same underlying distribution that results from selection, and that there is very little, if any, variability in selection across-trials.

We can also evaluate the extent to which guess 1 and guess 2 follow the same distribution by assessing quantile-quantile (QQ) plots of the observed conditional distribution and the distribution predicted by the model describing guess 2 as another independent sample from the same distribution as guess 1. If there is a shift in the distribution of guess 2 reports toward the serial position of guess 1 report, we should see an offset in the QQ plots around the target (0; shown in Figure 10a). In contrast, the only deviations from a diagonal line we see occurs in the tails, where random uniform guessing causes non-systematic deviations (Figure 10b).

These results further support our finding that guesses 1 and 2 are independent and identically distributed, indicating that responses are samples from the same underlying representation.

To sum up these results, in Experiment 1 we found that guess 1 and guess 2 on a given trial tend to come from adjacent serial positions, indicating that selective attention in time selects multiple letters on a given trial (thus ruling out the single-item selection hypothesis). Second, we found that guess 1 and guess 2 are independent, indicating that there is no shared across-trial variance between the two guesses, this rules out the contiguous all-or-none selection hypothesis and the contiguous graded selection hypothesis with any substantial amount of across-trial variability. Finally, we also found that the conditional guess 2 report distributions follow the predictions

of a model of guess 2 reports as another sample from the distribution of guess 1 reports; thus it seems that guess 1 and guess 2 are identically distributed. All together, these results support the hypothesis that on any given trial, attention selects a range of letters in a graded fashion, and the position of this selection window does not vary trial to trial. Responses have the statistical properties of independent and identically distributed samples from the graded selection distribution. A parsimonious account of these results describes selection as representing the uncertainty inherent in the inference about co-occurrence (the computational problem of the task) as a probability distribution over letters, from which responses are sampled.

3.4 Experiment 2

We have shown that selection in time (temporal selection) can be best described as contiguous graded selection with no detectable across-trial variability. In Experiment 2, we tested whether spatial selection also has the same properties. To do so, we employ a paradigm that exchanges the roles of spatial and temporal dimensions of the RSVP experiment to create conditions that are comparable to RSVP, but in the spatial domain. Specifically, in RSVP we display many letters in one location, separated in time: in Experiment 2, we display the same number of letters, at one point in time, separated in space. Thus, this design is similar to many historic iconic memory experiments (Averbach & Coriell, 1961).

3.4.1 Method

Participants. Eleven subjects from the Massachusetts Institute of Technology subject pool were recruited to participate. Subjects were between 18 and 35 years of age and were paid $10 for participation.

Materials. On each trial, subjects saw the 26 English letters presented simultaneously in a circle in a random arrangement. Each letter subtended approximately 2 degrees of visual angle, and the circle perimeter was at 6 degrees eccentricity. A line extending from fixation to the cued location served as the target cue. The cued location could be one of 13 points along the circle of letters (20 to 353 degrees in the monitor plane, separated in steps of 27 degrees). All display items were white on a black background, letters were in capitalized Courier font (Figure 11).

Each trial began with 1.5 s of fixation, then the cue was presented for 50 msec, followed by the letter array for 100 msec, followed again by the cue for 100 msec (see Figure 9).

The experiment was programmed in PsychToolbox (Brainard, 1997) on Matlab 7 on a Windows XP computer.

3.4.2 Procedure

Each participant began the experiment with two practice trials; the results of these trials were discarded. Following the practice trials, participants completed 5 blocks

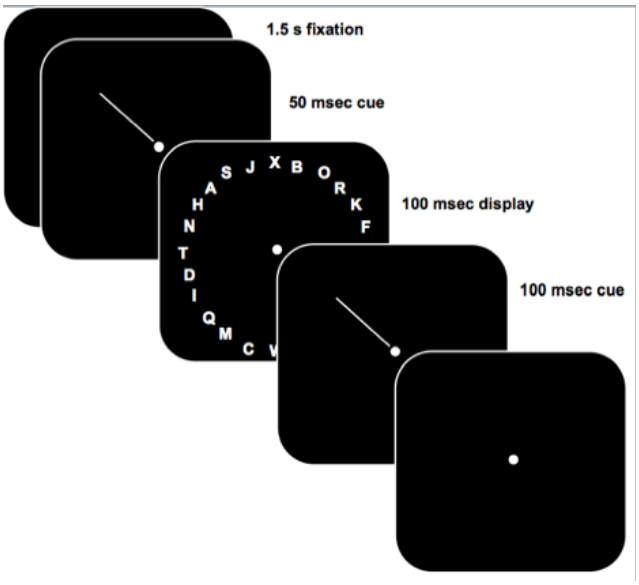

Figure 3-11: Experiment 2 design: A spatial version of RSVP, all letters are presented at the same point in time, spread across space, subjects must report the letter cued by the line, and are given four guesses to do so.

of 78 trials each. Each block contained 6 instances of each of the 13 possible cue locations, in a random order for each block.

At the end of each trial subjects were asked to make four guesses about which letter they thought was cued. Subjects reported the letters by pressing the corresponding keys on the keyboard. Just as in Experiment 1, duplicate letter reports were not accepted, and subjects were awarded 1, 0.5, 0.25, 0.125 points if they guessed the cued letter correctly on guesses 1-4, respectively. Again, as in Experiment 1, feedback and scoring reflected this instruction (in this experiment the average bonus was $1.60).

3.4.3 Results

Just as in Experiment 1, each letter appeared only in one (spatial) position on any given trial, thus we could identify the exact location where any given reported letter appeared. We could then compute the empirical histogram of reports around the cue across trials for any given guess. Figure 12a shows the empirical frequencies of reports for each guess and Figure 12b shows the logarithm of the ratio of observed to chance frequencies. Just as in Experiment 1, the histogram of reports across trials shows substantial variability, and again, above chance reports on the second guess (above 0 log observed-chance ratios; Figure 12b).

To determine if these results could arise from single item selection, or if multiple letters were selected on a given trial, we again analyzed the distribution of guess 2 reports around guess 1. As can be seen in Figure 13 the letters reported for guess 2 tend to be adjacent to the letter reported on guess 1 (for the 4 positions immediately adjacent to guess 1, guess 2 report frequency is above chance: all t values ¿ 4; df

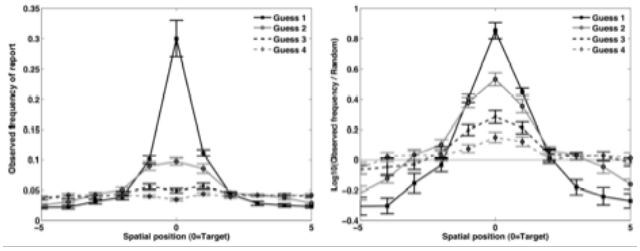

Figure 3-12: Experiment 2 results: the spatial distribution of reports for each of the four guesses subjects make on each trial. The x-axis corresponds to spatial position, where 0 is the target, negative positions are counterclockwise from the target, and positive positions are clockwise from the target. Red lines correspond to raw frequency data; blue lines are simulated chance performance (simulated given the condition that each item could only be reported once per trial); green lines are the logarithm of the ratio of observed to chance report frequencies. Error bars are 1 s.e.m.

= 8; ps¡.01). This indicates that in space, just as in time, selective attention selects several letters on a given trial.

We used the same logic as in Figure 8 of Experiment 1 to test whether the selected letters are selected in an all-or-none or graded fashion. If they were selected in an all-or-none fashion, then across-trial variability (translation of the selection window) is required to produce the observed graded across-trial histograms. Thus, again, we looked at the distributions of Guess 2 reports conditioned on different Guess 1 reports. Figure 14 shows the results of this analysis. Just as in the temporal case, in the spatial case the distribution of guess 2 reports does not depend on which item was reported on guess 1. We again compare the average reported position in the range of -1 through 1 when guess 1 came from serial position -2 and when it came from 2. We find no significant difference (p = .86, 95% confidence intervals on the mean shift are -0.12 to 0.10). As in experiment, we can also assess the independence of guess 1 and guess 2 by analyzing the correlation between guess 1 and guess 2 reports (using the same corrections as described in experiment 1). Again, in the spatial-selection case, just as in the temporal-selection case, we find no significant correlation (95% confidence intervals on the correlation coefficient are between 0.02 and 0.07, with an average of 102 trials included per subject). We can again assess whether the conditional distributions are identical by testing if there are any significant differences in the frequency of any reported spatial positions within the range of -2 to 2 for each of the conditional distributions. To this end we ran 30 pairwise comparisons, as in Experiment 1; although four were significant, none survived a Dunn-Sidak multiple-comparisons correction. Just as in experiment 1, the three analyses above indicate that guess 1 and guess 2 are independent, in that there is no evidence for any shared across-trial variance.

As in Experiment 1, we evaluate whether guess 1 and guess 2 are identically distributed by assessing whether conditional guess 2 reports follow the same distribution as would be predicted by a model that describes guess 2 as another sample

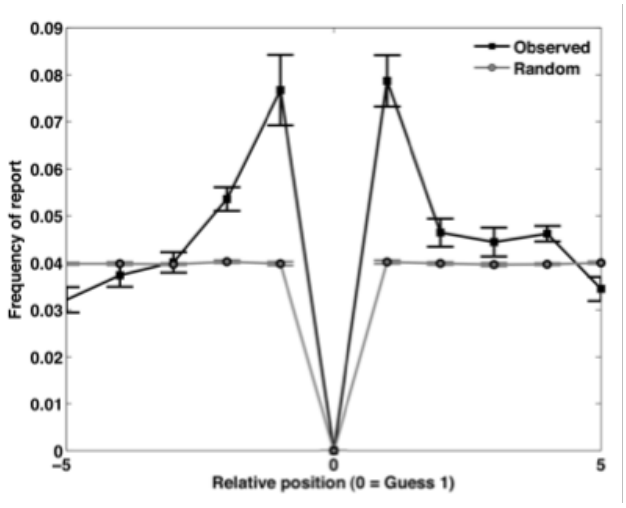

Figure 3-13: Experiment 2 results: Guess 2 reports around the position of the item guessed on Guess 1. Subjects are likely to report two spatially adjacent letters on guesses 1 and 2, indicating that multiple letters are selected on any given trial.

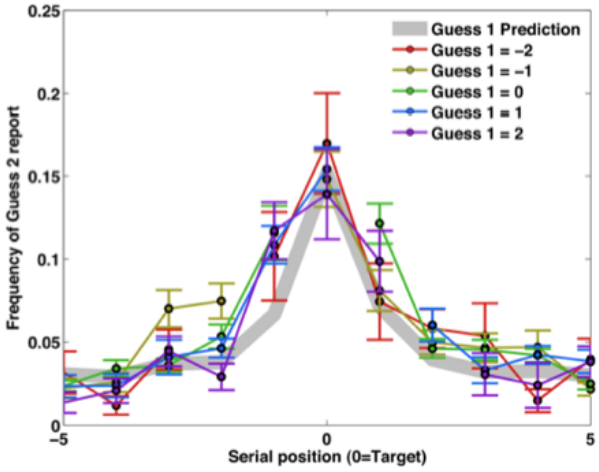

Figure 3-14: Experiment 2 results: conditional guess 2 reports as a function of spatial position.

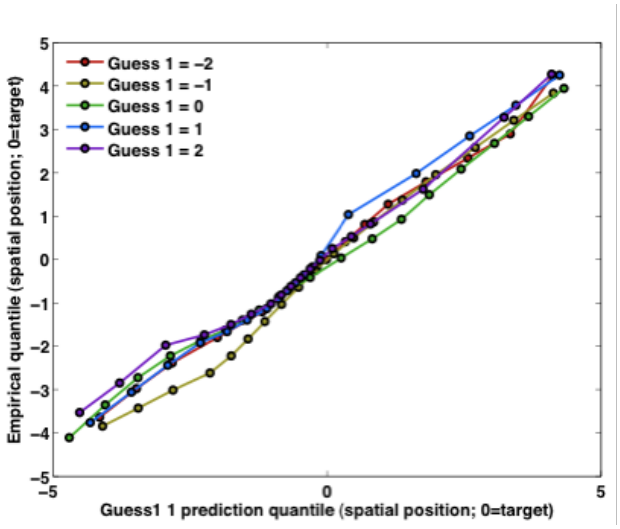

Figure 3-15: Experiment 2 results: QQ plots for predicted conditional guess 2 distributions and observed conditional guess 2 distributions.

from the guess 1 distribution (modulo increased random guessing and the fact that the same letter cannot be reported twice). The correlation between the model prediction (shown in Figure 14) and the observed conditional report frequencies is very high (r2=0.88, p¡.00001). Finally, we can again assess the quantile-quantile plots for the predicted distribution and the observed distributions (for a prediction of what a non-independent QQ plot would look like, see Figure 10a). Figure 15 shows these QQ plots: again, the only observable deviation from a diagonal occurs in the noisy tails, but not around the target (the point with highest probability), indicating that the predicted and observed probabilities match very well.

Again, in the spatial case, just as the temporal case, we see that selective attention selects a number of letters to varying degrees on any given trial, reflecting the uncertainty inherent in the task. This conclusion in the spatial case is reminiscent of the crowding phenomenon (He et al., 1996; Pelli et al., 2004): people are worse at identifying a cued letter when other, crowding, letters are nearby. Our data show that in such circumstances, attention selects multiple adjacent letters, and the actual reported letter is a sample from this selected distribution. Our findings are consistent with accounts of crowding as a limit in the spatial precision of selective attention (He et al., 1996). However, for our purposes, in spatial selection, just as in temporal selection, multiple responses on a single trial have the statistical properties of independent, identically distributed samples from an internal probability distribution that reflects the uncertainty inherent in the task.

3.5 Discussion

In two experiments we tested the mechanisms of visual selective attention. Specifically, we asked whether multiple items are selected to different degrees on each trial, as predicted by Bayesian models of cognition in which mental representations consist of multiple simultaneous hypotheses, each with a different graded probability of being true. The fact that many studies have reported graded distributions of responses in the average over many trials does not answer this question because such distributions could arise either from selection of multiple items on each trial, or from selection of a single discrete item on each trial, with some variability in the locus of selection across trials. To unconfound these two possibilities, subjects made multiple responses on each trial. In Experiment 1 we found that the temporal positions of intrusions from two guesses on one trial are uncorrelated. Because there was no correlation between errors on one trial, there is no shared spatial, or temporal, error between these two guesses. This observation means that there is no across-trial variance (or noise) in which items are selected, and therefore most of the variance seen in the final distribution of reports must occur within a given trial. Evidence of substantial within-trial variability indicates that subjects select a contiguous range of letters to varying degrees on every trial; thus, selective attention produces a representation equivalent to an internal probability distribution over likely targets (right panel of Figure 3). In Experiment 2 we extended these results to the domain of spatial selection. There too, our data indicate that selection creates a graded probability distribution over a range of possible targets, and subjects make responses by sampling guesses from this distribution. Thus, it seems that errors in visual selective attention tasks arise due to a process of sampling responses from internal representations that reflect the uncertainty inherent in the task.

Our results connect to three other lines of research. First, Sergent and Dehaene (2004) assessed whether conscious access is discrete or continuous using an Attentional Blink (Raymond et al., 1992) paradigm: when two targets in an RSVP stream appear in close temporal proximity, the second target is often missed due to failures of attentional selection (Vul, Nieuwestein, & Kanwisher, 2008). Sergent and Dehaene (2004) asked subjects to report the visibility of the second target with a continuous scale, and found that subjects used the scale in an all-or-none fashion: they reported either seeing or not seeing the target, without using any settings in between, suggesting that the target letter was not selected in a graded fashion. Our results suggest that subjects are not aware of the degree to which a given item was selected (and thus cannot choose the most likely alternative), but instead they must sample alternatives for report. Thus, it appears that while selective attention operates continuously, we are only aware of discrete samples from our internal probabilistic states, indicating that conscious access is discrete, as Sergent and Dehaene claim.

Second, the difference between continuous and graded selection and discrete conscious access bears on Boolean Map Theory (Huang & Pashler, 2007). Huang and Pashler describe a series of elegant experiments that suggest that subjects can select regions of space only via Boolean Maps a region of space may be selected, or not selected, with no states in between. However, evidence for this claim comes from

experiments that measure conscious access to the products of selection (e.g., mental rotation or transformation). There is no disagreement between our findings and those of Huang and Pashler: on our view, selection does not operate discretely, but rather continuously, selecting regions of space to varying degrees. However, access is discrete, and reflects a sample from the selected distribution. Thus, continuous selection and discrete access are not in opposition if access is limited to a discrete sample from the selected distribution.

Our conclusions are also consistent with a third line of research: Shih and Sperlings proposed account of visual selective attention as a spatiotemporal gate (2002). This account can be seen as an algorithmic-level analysis where we have offered an account at the level of computational theory (Marr, 1982). Our analysis of the computational problem entailed in selective attention tasks under uncertainty (detecting co-occurrences between cues and targets distributed over space or time) yields the same operations that Shih and Sperlings algorithmic level account proposed. The attentional gate proposed in their algorithm fulfills the computational role of uncertainty in the position of the cue. What Shih and Sperling refer to as spatio-temporal interference between items (interpretable as the persistence and point-spread functions of iconic memory), is computationally equivalent to what we refer to as the uncertainty about the spatiotemporal position of each letter. The process of multiplying the attentional gate function with the activation function of each letter and integrating over time, is the same computation one would undertake to perform the appropriate inference about co-occurrence. In Shih and Sperlings algorithm, the result of this multiplication and integration produces activation strengths in short-term memory these are computationally equivalent to a scaled probability distribution. Finally, the operation Shih and Sperling propose of adding noise to this short-term memory strength and responding by taking the maximally activated letter, may be equivalent to random sampling from a probability distribution (given certain conditions on the exact distribution of noise, e.g., such as variance scaling proportional to the activation strength (Ma et al., 2006). In short, the theoretical analysis of selective attention tasks that motivated our experiments is computationally isomorphic with the linear-systems account proposed by Shih and Sperling.

Several alternative accounts of our data cannot be ruled out by the present experiments. First, it could be the case that on each trial, multiple selection episodes are operating independently, each selecting one letter from a region around the target. We cannot rule out this account, as it could predict any pattern of data. However, on this account, an individual selection episode acts as the sampling process that we ascribe to post-selective process of retrieval from short-term memory; thus, instead of a probabilistic representation of the selected letters, as we advocate, this account must pose a probabilistic tendency to deploy selection episodes. Another alternative account is that there is complete certainty in the location of the cue, but there is substantial noise, or uncertainty, in the location of individual items, which are then coded with respect to their distance from the cue, and reported accordingly. Both of these accounts are plausible alternatives that should be addressed in future work. Tentatively we can say that other data from our lab (in which people are asked to report multiple features of one item) rules out the simplest version of this account

(Vul & Rich, in press). In general, completely ruling out "noise" in favor of "intrinsic uncertainty" as the source of variability in responses is impossible, as noise can be postulated to arise at any point in an arbitrarily complicated process model, thus making it consistent with just about any pattern of data. In our case, we think we have ruled out some intuitively simple accounts of noise in attentional selection, thus supporting the idea that in such tasks, intrinsic uncertainty coupled with a post-selection sampling process are responsible for variability in subjects responses.

There is an interesting tension in our data: we conclude that the gradation in the tendency to report a particular item reflects gradation in the degree to which each item is selected, rather than the average, across trials, of a set of all or none selection episodes of different items. We argue that this gradation reflects the result of uncertain inference about which item co-occurred with the cue. Thus, we postulate that the system represents uncertainty about where and when each item, and each cue, occurred. Usually, this uncertainty is purported to arise from noise that perturbs these measurements. However, we show no evidence of across-trial noise perturbing the spatiotemporal position of the cue (which would arise in translation of selection across trials); so why would there be uncertainty? This tension may be reconciled by supposing that the human visual system, through years of experience, has learned the amount of noise perturbing its measurements of the world, and the learned uncertainty persists in controlled laboratory settings when actual noise is eliminated through precise digital presentation. If so, we predict that the uncertainty in selection (as measured by spatiotemporal variability of reported items) would decrease with sufficient training in laboratory or video-game settings this is a promising direction for future research.

In sum, our results provide evidence of a sampling process that connects graded internal probability distributions with discrete conscious access and responses. These results dovetail with findings from a very different domain: when subjects are asked to guess arbitrary facts about the world (e.g., "What proportion of the worlds airports are in the US?") multiple guesses from one subject contain independent error thus, averaging two responses from one subject produces a more accurate guess than either response alone a "crowd within" (Vul & Pashler, 2008). The "crowd within" results, and the results in this paper are both predicted by the idea that the mind operates via probabilistic inference (Chater et al., 2006), and solves complicated probabilistic inference problems by sampling (Vul et al., n.d.). Internal representations are graded probability distributions, yet responses about, and conscious access to, these representations is limited to discrete samples. Our mind appears to perform Bayesian inference without our knowing it.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Independent sampling of features enables conscious perception of bound objects

4.1 Thesis framing

In the previous chapter, I showed that multiple guesses about a single feature in a visual selective attention task contain independent spatial or temporal error. Does this "sampling" behavior in visual selection help make sense of any attentional phenomena? In this article with Anina Rich, I explore the implications of the spatiotemporal sampling hypothesis for the "binding problem" (Treisman & Gelade, 1980).

This chapter was published as: (Vul & Rich, in press)

4.2 Abstract

Decades of research suggest that selective attention is critical for binding the features of objects together for conscious perception. A fundamental question, however, remains unresolved: How do we perceive objects, albeit with binding errors (illusory conjunctions) when attentional resolution is poor? Here we use a novel technique to ask how features are selected to create percepts of bound objects. We measured the correlation of errors (intrusions) in color and identity reports in spatial and temporal selection tasks under conditions of varying spatial or temporal uncertainty. Our findings suggest that attention selects different features by randomly sampling from a probability distribution over space or time. Thus, veridical perception of bound object features arises only when attentional selection is sufficiently precise that the independently sampled features originate from a single object.

4.3 Introduction

We effortlessly perceive scenes comprised of numerous objects with many varied features. We perceive the correct combination of colors, shapes, and motion that make up these objects (e.g., a red car driving north, a blue bicycle going south). Each object is seen as a cohesive whole, despite the fact that different features (e.g., color, shape, motion) are processed in anatomically segregated parts of the brain (Livingstone & Hubel, 1988).

Although usually we successfully perceive the correct conjunctions, when selective attention is diverted or impaired, binding of object features can go awry, causing illusory conjunctions of incorrect features (e.g., an object with the color of one item and the shape of another; Robertson, 2003; Treisman & Gelade, 1980). Such illusory conjunctions highlight the challenge known as the binding problem (Wolfe & Cave, 1999): How does the brain combine information from different specialized areas to provide our subjective experience of cohesive objects? Although psychophysical and physiological evidence suggests that conjunctions are represented in primary visual cortex (Sincich & Horton, 2005) and are formed without attention (Humphrey & Goodale, 1998) or consciousness (Vul & MacLeod, 2006), our conscious perception of objects seems to require feature binding by attention (Treisman, 2006).

Most proposals about how attention binds features together for conscious perception suggest that we infer which features belong to one object by virtue of their location. Feature Integration Theory posits that visual attention conjoins features into object files (Treisman & Gelade, 1980; Treisman & Schmidt, 1982), by directing an attentional spotlight to a spatial location and selecting the features therein. Boolean Map Theory proposes that perception is mediated by a map that defines locations as either selected or not, and features within the same map are bound together (Huang & Pashler, 2007; Huang et al., 2007). However, these accounts leave a fundamental question unanswered: when the attended location is not precise enough to encompass only one object, how are features selected for conscious perception?

Outside the attended region, multiple features seem to be aggregated though a process of statistical summary (Chong & Treisman, 2005; Alvarez & Oliva, 2008, 2009); however, this process produces averages of features rather than illusory conjunctions. Therefore, some suggest that individual features within an attended region are randomly chosen for perception (Ashby, Prinzmetal, Ivry, & Maddox, 1996; Huang & Pashler, 2007; Treisman & Schmidt, 1982; Vul et al., 2009). On this account, illusory conjunctions arise because different features (e.g., color and form) are chosen independently (Ashby et al., 1996). This account predicts independent intrusions from different features; for example, the report of a color from one item would not predict a report of the form from the same item. Unfortunately, attempts to demonstrate the independence of feature intrusions using accuracy measures (Bundasen, Kyllingsbaek, & Larsen, 2003; Isenberg, Nissen, & Marchak, 1990; Nissen, 1985) have been controversial because irrelevant task factors could introduce, or eliminate, dependence in the accuracy of different feature reports (Monheit & Johnson, 1994; Wolfe & Cave, 1999; e.g., trials where subjects blinked and missed all the features would induce dependence in accuracy between reports of two features). Such extraneous sources of

dependence must be factored out to assess independence of feature binding.

Here we adopt a general statistical framing of the binding process. Like previous accounts, we assume that subjects assess object features by estimating the location of the object, and evaluating which features were present in that location. Since there will be some uncertainty in location estimates of both the object and the features, this process amounts to probabilistic inference about the co-occurrence of features with the location of interest (Ashby et al., 1996; Vul et al., 2009). In a scene with multiple objects, attention mediates this inference by establishing the location of the relevant object in space and time, thus creating a probability distribution that describes the estimated spatiotemporal location of interest. The claim that features are randomly chosen from within a selected region postulates that features are sampled from this probability distribution. On this account, decreased precision of attention amounts to worse estimates – and thus increased uncertainty – about the location of the object. In turn, greater uncertainty about the location of the object will result in greater errors in feature intrusions.

Given this statistical framing, we designed a new measure to directly test whether conscious perception of conjunctions is comprised of features independently sampled from a probability distribution over space or time. Our goal is to assess whether feature intrusions for different feature dimensions (e.g., color and form) are uncorrelated, as they would be if independently sampled given location uncertainty. In contrast, if feature intrusions are correlated, this suggests that they share a source of error, such as internal noise in the location of the attended region. Instead of looking at the accuracy of different feature reports, we evaluate the spatial positions of the reported features, and ask whether these are correlated between different feature dimensions. When subjects are asked to report both the color and identity of a letter cued in space, they do not always report the correct color and letter: subjects frequently report the spatially proximal colors and letters. Our question is: if a subject reports the color to the right of the target, does this also predict that the subject will report the letter to the right of the target? We can answer this question by looking at the correlation in spatial position errors between the two features. Using this measure, we can detect systematic relationships between feature intrusions, and can therefore assess the dependence of one feature report on the other while factoring out shared task factors.

We can ensure that independence, as measured by a lack of correlation, is not due to limitations of memory (Tsal, 1989; Wolfe & Cave, 1999), task demands, or statistical power, by introducing external noise to the cue to simulate the possible effects of internal noise. In this manipulation, the cue is less accurate, effectively pointing to items on either side of the target on some trials. This should cause a systematic relationship in the errors of color and letter identity, because errors in cue position will contribute to the position error of both feature reports. We can therefore verify that our method is able to detect correlations of position intrusions across features when we know that they should be present.

Our results show that in both space (Experiment 1) and time (Experiment 3), illusory conjunctions arise from a process that samples features independently there is no correlation between intrusions in color and intrusions in letter identity. Furthermore,

this lack of correlation cannot be ascribed to limitations of memory, task demands, or statistical power, because, in both space (Experiment 2) and time (Experiment 4), the external noise manipulation produces reliably correlated feature intrusions.

4.4 Experiments 1 and 2: Binding in space

4.4.1 Method

Participants. In Experiment 1 (spatial uncertainty), 10 participants (6 female; aged 18-40 years) from the Massachusetts Institute of Technology subject pool were paid for participation. In Experiment 2 (spatial noise), 12 participants (10 female; aged 18-40 years) from the Macquarie University subject pool were given course credit or paid for participation.

Materials and Design. Subjects viewed a brief presentation of 26 colored capital letters arranged in a circle equidistant from fixation, and reported the color and letter identity of one item cued as the target. The colors and letters were unique within the five items around the target, allowing us to identify on every trial the spatial or temporal position corresponding to each reported feature, relative to the cued item. The 26 English letters, were presented in a random order in Courier font around a (6 degree diameter) circle. Each letter was randomly assigned one of five colors with the constraint that the target letter was the center of a set of five uniquely colored letters. At a viewing distance of 50 cm, the letters subtended 1.3 degrees of visual angle at 6 degrees eccentricity. A white line extending from the centre of the display (4 degrees in length) cued the target.

In the spatial uncertainty condition (Experiment 1), we manipulated the information available about the cue location (and thus, the precision of attention) by accurately cueing the target location at a variable pre-cue interval before the onset of the letter array. The time between cue onset and the letter display onset (pre-cue time) was 0, 100, or 200 msec. These values were chosen to discourage saccades to the target location[1]. These three pre-cue conditions were randomly intermixed within a block. After the cue, the stimulus array was presented for 100 msec. Shorter pre-cue intervals provided less information about the cue direction, thus decreasing spatial precision in estimated target locations. Our key question is whether this imprecision is best described as uncertainty or as internal noise when evaluating the conjunctions of items.

The spatial noise condition (Experiment 2) was designed to illustrate the effects of shared noise on a given trial. For this, pre-cue time was fixed at the longest duration (200 msec), but we added spatial noise to the location indicated by the cue. The noise was Gaussian with a standard deviation equal to 1, 0.5, or 0 times the spacing of the letters (13.8, 6.9, and 0 degrees of arc, respectively). These noise mag-

[1]On each trial we asked subjects if they moved their eyes (to remind them to fixate) and discarded trials when they reported having done so. Importantly, eye-movements are not a major concern because saccades to the target location would induce (rather than mitigate) a correlation in feature reports.

nitudes approximated the standard deviation (in items) of responses in Experiment 1. On average across trials, the cue pointed to the correct target, but on any one trial, it could point slightly off-target. External noise in the cued location simulated the possible contribution of internal noise in attended locations to feature intrusions in Experiment 1. We expected that position intrusions would be correlated across features in this condition because, by design, both color and identity reports share a common source of error: the external noise in cue position. Thus, this manipulation tests whether we can detect a correlation between feature intrusions when we know it should be present.

4.4.2 Procedure

Participants completed 5 blocks of 60 trials each. Each trial began with a fixation cross for 500ms, then the cue line appeared, then the target display appeared after a variable (Experiment 1) or fixed (Experiment 2) interval. Participants used the number keys to indicate the target letter identity from five options, and the target color from five options in a separate display. Subjects were awarded one point for each feature correctly reported, leading to a score of 0, 1 or 2 per trial. The target location, color-letter pairings, the position of each colored letter in the display, and the report order (color or identity first) were randomly chosen on each trial.

4.4.3 Results

Since the colors and letters were unique within the five items around the target, we could identify the spatial position corresponding to each reported feature, relative to the cued item. We used this information to construct the joint distribution of color and letter reports how often each of the 25 (5 letters by 5 colors) logically possible conjunctions was reported.

We quantified the spatial error of a given feature report by its spatial deviation: reporting the feature of the target has a deviation of 0; reporting the feature of the item next to the target has a deviation of +1; and two items away will be +2. We (arbitrarily) labeled clockwise deviations as positive, and counter-clockwise deviations as negative. Thus if the cued item is the red L in Figure 1a, a report of yellow reflects a spatial deviation in color report of -1 , a report of W would also be a -1 intrusion, whereas reporting K would be a +2 intrusion. The exact scoring is unimportant, what is critical is that we can calculate the magnitude and direction of spatial position intrusions in both color and letter reports.

The variance of the spatial deviations describes the imprecision of feature reports. If this imprecision arises from independent sampling of color and letter identity given some spatial uncertainty about the location of the target, there should be no correlation between features. In contrast, if this imprecision reflects internal noise in the estimated target location, then this noise will contribute to both color and letter errors, resulting in a correlation in their spatial intrusions . The covariance of colorreport deviations and letter-identity-report deviations is a direct measure of the independence of feature intrusions. We measure the correlation of intrusions via their

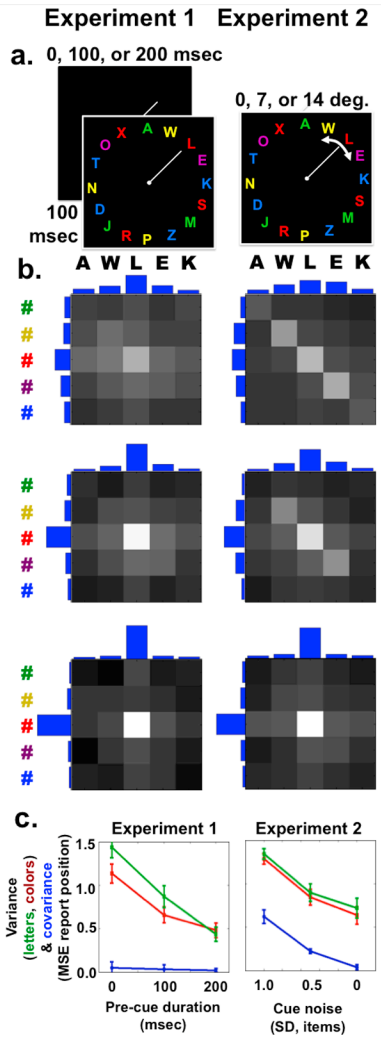

Figure 4-1: Results for Experiments 1 and 2: Binding in space. (a) An array of letters was presented in a ring around fixation for 100 msec. The target was cued for report by a white line. We varied the spatial uncertainty by manipulating the time between onset of the cue, and onset of the letter array (pre-cue time; Expt.1, left panel). We varied external noise by perturbing the angular direction of the cue with Gaussian noise (Expt. 2, right panel). (b) Joint frequency histograms denoting the report frequency of each of the possible conjunctions for the uncertainty manipulation (left panels; 0 msec, 100 msec and 200 msec pre-cue condition, respectively), and the noise manipulation (right panels; 200 msec pre-cue, 1.0, 0.5 and 0 noise conditions, respectively), as a function of increasing report variance. Increasing lightness reflects greater frequency of report. Marginal error histograms (frequency of report of each feature) for color and letter intrusions are shown in blue. (c) Variance and covariance (mean squared intrusion position) for the joint report distributions. Increasing both cue uncertainty and cue noise increased the variance for reporting letters (green) and colors (red). However, only increasing noise altered the covariance (blue) of letter and color reports.

covariance (an unnormalized measure of the correlation [2]) because the units of covariance and variance are directly comparable. With this measure, we can test directly whether a spatial intrusion in color predicts an intrusion of letter identity and vice versa. A systematic relationship between errors on the two features will be detected as non-zero covariance in the joint distribution of errors – this covariance is our measure of the (in)dependence of letter and color reports.

Figure 1b (left panels) shows the joint report distributions for the different pre-cue durations (0ms, 100ms, 200ms) for Experiment 1. Participants reported the correct conjunction (central item) more often (lighter squares, central blue bars in the individual feature histograms) as the pre-cueing time increased. This demonstrates that the pre-cue duration successfully manipulated the spatial precision of attentional selection: the spatial variance of intrusions was lower with longer pre-cue durations (Figure 1c; 95% confidence intervals (CIs) on the slope of the decrease [-6.5 -3.5] for letter reports and [-4.6 -1.9] for color reports; R2=0.62 and 0.47, respectively; slopes are defined as the change in variance as a function of pre-cue duration in seconds). Thus, with a longer cue exposure – and therefore more information about its spatial position – inference about which item was cued becomes more precise (or, alternatively, a narrower spatial window is selected around the target), increasing accuracy and precision of both feature reports.

Critically, across the large variations in performance in the two feature dimensions, pre-cueing never increased the covariance of the feature intrusions: it remained at zero for all conditions (Figure 1c, left panel; 95% CIs: 0 msec pre-cue [-0.1 0.19]; 100 msec pre-cue [-0.08 0.14]; 200 msec pre-cue [-0.03 0.06]). None of the pre-cue conditions induced dependence between color and letter intrusions feature intrusions were statistically equivalent to independent, identically distributed samples drawn from a probability distribution over space.

Previous work investigating the independence of feature reports suffered from an inability to distinguish visual selection rather than memory as the source of unbinding (and independence) in feature errors (Tsal, 1989). Here, this is less of a problem, since we look at spatial correlation rather than simply accuracy. Nonetheless, we can further demonstrate the source of errors by contrasting the effects of spatial uncertainty about the cue location with the effects of external noise.

In Experiment 2, we presented the easiest pre-cue condition (200 msec), but perturbed the spatial position of the cue by adding external noise to it that was matched to the variability observed in Experiment 1 thus, the cue did not always point to the correct target (see Methods; Figure 1a, right panel). Effectively, this makes the cue a less accurate indicator of target location. Importantly, because the error in the cue position will affect both color and letter reports, this external noise should introduce a correlation between the feature reports. Thus, this condition verifies that our method can detect a correlation when both features do in fact share a common source of error.

The results of Experiment 2 are shown in Figure 1b and 1c (right panels). As

[2] $r_{xy} = \sigma_{xy}/(\sigma_x\sigma_y)$ where r_{xy} is the correlation, σ_{xy} is the covariance, and σ_x, σ_y are the marginal standard deviations of x and y.

expected, adding noise to the angular position of the cue decreased accuracy (the spatial variance of feature intrusions increased: 95% CIs and R^2 on the slope of this increase: for color [0.42 0.89], $R^2 = 0.48$; for letter identity [0.36 0.89], $R^2 = 0.41$). More importantly, cue noise added correlated variance: the covariance also increased (Figure 1c, right panel; 95% CIs: 0 noise [-0.03 0.1], 7 degree noise [0.17 0.28], 14 degree noise [0.46 0.78]). This indicates that subjects are able to report correlated features, and such correlations are detectable using our methodology. Thus, the independence we observed in Experiment 1 does not arise due to unbinding in memory or limited statistical power, but rather due to independent sampling of features given spatial uncertainty in attentional selection.

4.5 Experiments 3 and 4: Binding in time

Although binding in space is the canonical form of feature binding, illusory conjunctions and mis-bindings occur in time as well (Botella, Arend, & Suero, 2004; Botella, Suero, & Barriopedro, 2004). If random sampling of features is a general mechanism of visual attention (Vul et al., 2009), our results should replicate in the temporal domain. To test this, we rearranged our 26 colored letters in a rapid serial visual presentation (RSVP) at fixation. One letter was cued by an annulus that appeared simultaneously with the letter. We manipulated the temporal uncertainty (precision of attentional selection in time) by varying the presentation rate (Experiment 3, Figure 2a). Again, we contrasted this manipulation of uncertainty with a manipulation of external noise (Experiment 4).

4.5.1 Method

Participants. Participants were drawn from the Macquarie University subject pool and received course credit or were paid. There were 14 participants (10 female; aged 18-45 years) in Experiment 3 (temporal uncertainty), and 12 participants (10 female; aged 18-45 years) in Experiment 4 (temporal noise).

Materials. A rapid stream of capitalized Courier font colored letters was presented at fixation, with a white ring cueing the target. With our resolution (1024x768), monitor (Dell P992), and viewing distance (57 cm), each letter subtended 2.9 degrees of visual angle.

In the temporal uncertainty condition (Experiment 3) the presentation rate was 13.3, 10 or 6.7 items/second on randomly intermingled trials. The item/inter-item-blank durations were 45/30, 60/40, and 90/60 msec, respectively. Cues appeared concurrently with the target.

In the external noise condition (Experiment 4), the presentation rate was fixed at 6.7 items/second, but we added temporal noise to the cue onset. The noise was Gaussian with a standard deviation equal to 1.5, 0.8, 0 times the item presentation time (225, 120, and 0 msec). Thus, in the non-zero noise conditions, the cue onset could be slightly earlier or later than the target (even during items preceding the target).

4.5.2 Procedure

Participants completed 5 blocks of 78 trials. As with the spatial experiments, subjects were awarded points for correctly reporting the color and letter identity of the target on each trial. The target onset (item 6 to item 20 of 26), letter order, color order, report order, and presentation rate (Experiment 3) or cue noise (Experiment 4) conditions were all randomly chosen on each trial.

4.5.3 Results

For binding in time, just as in space, increasing temporal uncertainty of selection in this case, by accelerating the RSVP stream decreased accuracy. This is evident in the joint and marginal report histograms (Figure 2b; left panel) and can be quantified by the variance of the temporal position deviations of the reported features (Figure 2c, left panel). Variance of temporal intrusions increased with RSVP rate (95% CIs for the slope of this increase and R^2: for color [-10.2 -5.0], $R^2 = 0.48$; for letter identity [-12.0 -8.1], $R^2 = 0.73$). Just as in our manipulation of spatial uncertainty, manipulations of temporal uncertainty had virtually no effect on the dependence of feature intrusions (the covariance of intrusion positions did not correlate with RSVP rate: 95% CI on the slope: [-1.1 0.1]; $R^2 = 0.076$). Although neither the fastest nor the slowest RSVP rate showed a covariance significantly greater than zero (95% CIs of [-0.01 0.09] and [-0.01 0.02], respectively), the medium rate had a significantly non-zero covariance (95% CI: [0.03 0.07]). However, compared to the overall variance of temporal intrusions, the magnitude of this non-zero covariance is negligible.

In contrast to temporal uncertainty, the addition of external noise to the temporal position of the cue (Experiment 4; Figure 2, right panels), increased both the variance of reports (Figure 2c, right panel: 95% CI for slope and R^2: for color [0.61 0.89], $R^2 = 0.78$; for letter identity [0.54 0.8], $R^2 = 0.76$), and the covariance of feature intrusions (95% CI [0.52 0.72], $R^2 = 0.83$).

The trend towards some non-zero covariance in the case of temporal uncertainty raises the possibility that there may be some information about feature conjunctions in one of our attention conditions. We hesitate to make this conclusion, however, because the effect is so small compared with the fluctuations in intrusion variance: the correlation between color and letter intrusions accounted for an average of 8% of the position errors, and the magnitude of the covariance did not change with temporal uncertainty (unlike variance). In contrast, when we manipulated temporal noise, covariance changed with variance, and accounted for 43% of the variability in letter and color intrusions. Thus, we conclude that in time, just as in space, feature reports are well described as statistically independent samples from a spatiotemporal probability distribution.

4.6 Discussion

Attentional feature binding appears fundamental to our conscious visual experience, allowing us to effortlessly perceive objects as cohesive structures of different features.

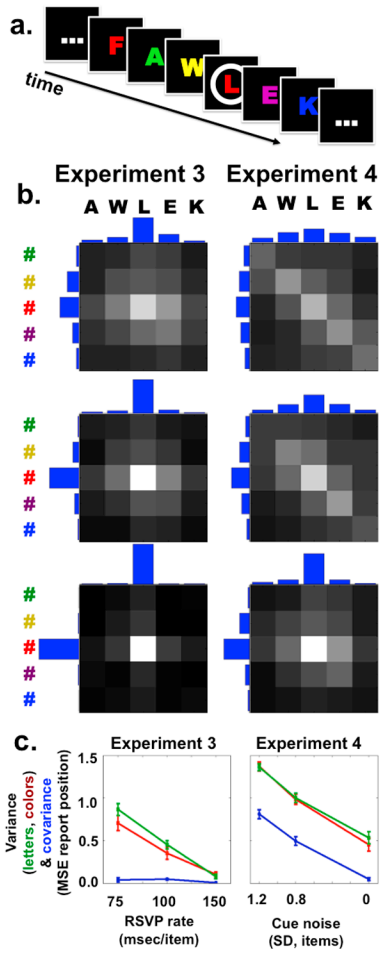

Figure 4-2: Results for Experiments 3 and 4: Binding in time. (a) Letters were presented in a rapid sequence at fixation. The target was cued for report by a white annulus. We varied the uncertainty (Exp. 3) inherent in cueing by manipulating the rate at which items appeared (RSVP rate). We varied external noise (Exp. 4) by perturbing the onset time at which the cue appeared with Gaussian noise. (b) Joint frequency histograms with which every possible conjunction was reported for the (left) uncertainty manipulation and the (right) noise manipulation. (c) Increasing both cue uncertainty and cue noise increased the variance for reporting letters (green) and colors (red). However, only increasing noise altered the covariance (blue) of letter and color reports.

Here, we examined how selective attention achieves this critical process by asking what determines the perceived features of an object when attention fails to create the veridical conjunction. Our results show that two features perceived as a conjunction are statistically equivalent to two independent samples from a probability distribution (attentionally selected region) in both space and time. Accurate binding of features is therefore not a special mechanism or action by selective attention, but merely the limiting case when the attentionally selected region is narrow enough to encompass only one object on our account, this arises when there is sufficient information about the spatiotemporal location of the attended object.

There are a few alternative explanations about the source of feature intrusions in binding. First, there may be internal noise in estimating the spatiotemporal location of an object. That is, noise in the visual system results in selection of an incorrect location. Our data rule this out, however, because both the letter and the color report would share this noise, producing a correlation in their intrusions. Second, there may be independent spatiotemporal noise for different features. In this case, if participants make multiple guesses about a single feature (guess the cued letter; then make a second, different, guess), feature intrusions across the two guesses should be correlated. Recent data, however, shows that multiple guesses about one feature also contain independent error (Vul et al., 2009). Thus, the intrusions in our present results seem unlikely to arise from internal noise about feature locations.

Our interpretation of the current results is that the rapid presentation of a spatial or temporal cue provides insufficient information about the spatial or temporal position of the cued object. Participants are left with some uncertainty about the location of the relevant item; and the window of attentional selection corresponds to a probability distribution over space and time, describing the inherent uncertainty of the task. This spatiotemporal uncertainty yields a probability distribution over features likely to have been present in that location. Subjects then independently sample features from this probability distribution., Thus, the color and letter responses are sampled from a probability distribution that encompasses the likely target as well as the surrounding items. Crucially, the two features are sampled independently, as demonstrated by a lack of correlation in their errors. On this account, both veridical and illusory binding arise from the way visual attention copes with uncertainty: approximation through sampling (Vul et al., n.d., 2009).

These results connect binding to a growing literature that suggests that, in general, the human mind implements complex probabilistic computations via sampling, resulting in responses that appear to be probability-matched to beliefs (Vul et al., n.d.; Vul & Pashler, 2008; Vul et al., 2009; Goodman et al., 2008; Sanborn & Griffiths, 2008; Herrnstein, 1961). Our data suggest that visual attention acts as a sampling process to select visual features for our conscious perception, rather than completing a special binding process: Veridical binding is just the limiting case of this sampling process, when the spatiotemporal window from which features are independently sampled is narrow enough to contain only one object.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Measuring the Crowd Within: Probabilistic Representations Within Individuals

5.1 Thesis framing

How general is the sampling hypothesis. Thus far I have shown that errors and variation in visual attention have the statistical properties of independent samples from a probability distribution, but does this hold outside of attention in everyday tasks? Circumstantial evidence (e.g., Griffiths & Tenenbaum, 2006) suggests that this is the case, but this hypothesis has not been tested by directly evaluating multiple responses from one subject. In this chapter I will ask whether multiple guesses about real-world quantities from one subject contain independent error, thus yielding a "wisdom of crowds" (Suroweicki, 2004) benefit from a single person.

This chapter was published as: (Vul & Pashler, 2008)

5.2 Introduction

A crowd often possesses better information than the individuals that comprise it. For example, if people are asked to guess the weight of a prize-winning ox (Galton, 1907), the error of the average response is substantially smaller than the average error of individual estimates. This fact, which Galton interpreted as support for democratic governance, is responsible for the success of polling the audience in "Who Wants to be a Millionaire" (Suroweicki, 2004) and the superiority of combining financial forecasts (Clemen, 1989). This wisdom of crowds effect is now agreed to depend on a statistical fact: the crowd average will be more accurate so long as some of the error of one individual is statistically independent of the error of other individuals as it seems almost guaranteed to be.

What is not obvious a priori is whether a similar improvement can be obtained by averaging two estimates from a single individual. If one estimate represents the best information available to the person, as common intuition suggests, then a second guess

will simply add noise, and averaging the two will only decrease accuracy. Researchers have previously assumed this view and focused on improving the best estimate (Hirt & Markman, 1995; Mussweiler, Strack, & Pfeiffer, 2000; T. Stewart, 1999).

Alternatively, single estimates may represent samples drawn from an internal probability distribution, rather than a deterministic best guess. On this account, if the internal probability distribution is unbiased, the average of two estimates from one person will be more accurate than a single estimate. Ariely and colleagues (Ariely et al., 2000) predicted that such a benefit would exist when averaging probability judgments within one individual, but did not find evidence of such an effect. However, probability judgments are known to be biased toward extreme values (0 or 1), and averaging should not reduce the bias of estimates; instead, if guesses are sampled from an unbiased distribution, averaging should reduce error (variance) (Laplace, 1818; Wallsten, Budescu, Erev, & Diederich, 1997).

Probabilistic representations have been postulated in recent models of memory (Steyvers et al., 2006), perception (Kersten & Yuille, 2003), and neural coding (Ma et al., 2006). Naturally, it is consistent with such models that responses across people are distributed probabilistically, as the "wisdom of crowds" effect shows. However, there has been no evidence that within a given person knowledge is represented as a probability distribution. Finding any benefit of averaging two responses from one person would yield support for this theoretical construct.

5.3 Method

We recruited 428 participants from an internet-based subject pool. We asked participants eight questions probing real world knowledge (derived from the CIA World Factbook, e.g., "What percent of the world's airports are in the United States?"). Participants were asked to guess the correct answer. Half were unexpectedly asked to make a second, different guess regarding each question immediately upon completion of the questionnaire (immediate condition); the other half made a second guess 3 weeks later (delayed condition).

5.4 Results

The average of two guesses from one individual (within-person average) was more accurate (lower mean squared error) than either guess alone (see Figure 1a; immediate: average - guess1, $t(254)=2.25$, $p<.05$, average - guess2, $t(254) = 6.08$, $p<.01$; delayed: average - guess1, $t(172)=3.94$, $p<.01$, average - guess2, $t(172)=6.59$, $p<.01$), indicating that subjects do not produce a second guess by simply perturbing the first, but that error of the two guesses was somewhat independent. This averaging benefit cannot be attributed to subjects finding more information between guesses, because the second guess alone was less accurate than the first (Figure 1a; immediate: $t(254)=3.6$, $p<.01$; delayed: $t(172)=2.8$, $p<.01$). Moreover, the averaging benefit was greater when the second guess was delayed by three weeks (difference in error between

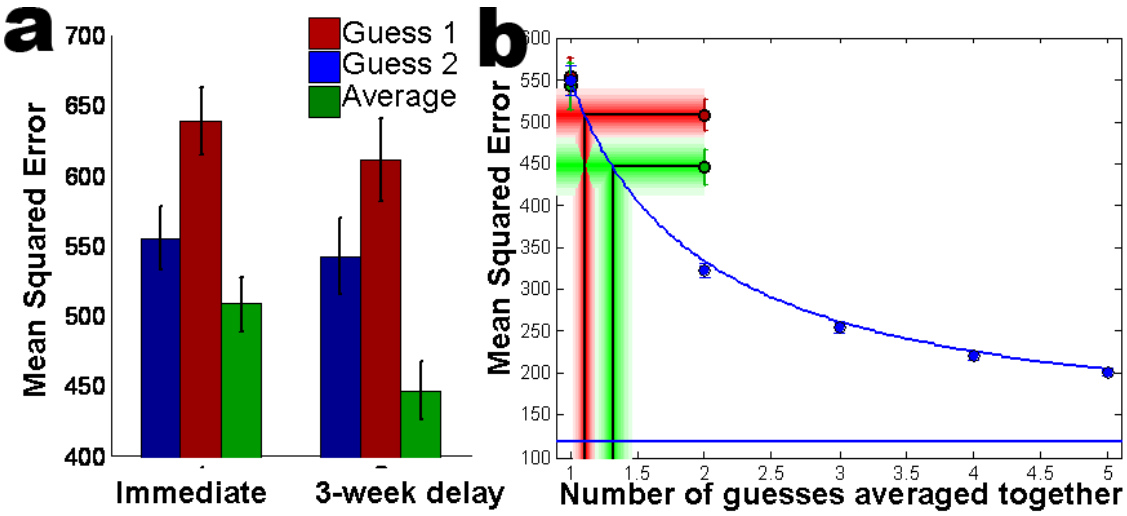

Figure 5-1: (a) Mean squared error (MSE) for the immediate and 3-week delayed second guess conditions. Guess 1 (blue) is more accurate than guess 2 (red) for both conditions, but the average of the two guesses (green) is more accurate than either guess alone. This benefit is greater in the delayed second guess condition. (b) Data points are MSE as a function of number of guesses averaged together, for guesses from independent subjects (blue); a single subject, immediate condition (red); and a single subject, 3-week delay (green). Blue curve shows convergence to the population bias (the error of the guess averaged across all people – horizontal blue line). Through interpolation (black lines) we compute that two guesses from one person are worth 1.11 guesses from independent people, if immediate, or 1.32 if delayed by 3 weeks. Shaded regions are boot-strapped confidence intervals up to 0.9. All differences are significant (p¡0.05). Error bars are 1 s.e.m.

the first guess and the average was greater in the delayed than the immediate condition: t(426)=2.12, p¡.05; 95% confidence intervals for percentage of error reduced: immediate: [2.5% 10.4%]; delayed: [11.6% to 20.4%]). Thus, one benefits from polling the crowd within, and the inner crowd grows more effective (independent) when more time elapses between guesses.

We can compare the efficacy of within-person averaging to across-person averaging, via hyperbolic interpolation (Figure 1b). According to the central limit theorem, if error across subjects is independent, the average of N guesses from N people should converge to the group bias proportionally to 1/N. This hyperbola fits the across-person averages perfectly (R2=1; Figure 1b). However, N guesses from one person are not as beneficial as N guesses from N people. The benefit of N guesses from one person can be described as $1/(1+\lambda(N-1))$, where λ is the proportion of a guess from another person that a guess from the same person is worth. λ can be estimated by interpolating the benefit of within-person averaging onto the hyperbola representing the benefit of across-person averaging. Thus, we compute the value of N (in guesses from multiple people) that two guesses from one person are worth. This value is 1.11

(λ = 0.11) for two immediate guesses, and 1.32 (λ =0.32) for two delayed guesses. Simply put, you can gain about one-tenth as much from asking yourself the same question twice as you can from getting a second opinion from someone else, but if you wait 3 weeks, the benefit of re-asking yourself the same question rises to 1/3 the value of the second opinion. One potential account for this immediacy cost is that subjects are biased by their first response to produce less independent samples (while a delay mitigates this anchoring effect).

5.5 Discussion

Although people assume that their first guess about a matter of fact exhausts the best information available to them, a forced second guess contributes additional information, such that the average of two guesses is better than either guess alone. This observed benefit of averaging multiple responses from the same person suggests that responses made by a subject are sampled from an internal probability distribution, rather than deterministically selected based on all the knowledge a subject may have.

Temporal separation of guesses increases the benefit of within-person averaging by increasing the independence of guesses, thus making another guess from the same person more like a guess from a completely different individual. Beyond theoretical implications about the probabilistic nature of knowledge, these results suggest a quantitative measure of the benefit of "sleeping on it".

Chapter 6

General Discussion: Towards a Bayesian cognitive architecture

There are two global challenges to extending rational analysis from the computational to the algorithmic level. The main body of this dissertation addressed the first question: How can ideal Bayesian inference be approximated in by the human mind given its limited cognitive resources? Thus far I have argued that just-in-time sampling algorithms are a likely candidate as a resolution to this problem, connecting ideal Bayesian inference at the computational level to the processing level of cognition.

However, a second challenge to connecting the computational and algorithmic levels of description remains: How should cognitive resources be treated within a rational analysis framework? A cognitive architecture describes the processes by which people utilize limited cognitive resources. However, our use of cognitive resources must depend on our goals: we must look at, attend to, compute, and remember the details relevant for our goals. In this discussion I will describe an outline for a Bayesian cognitive architecture that captures these ideas by casting the use of sensory or cognitive resources as internal actions, governed by statistical decision theory.

6.1 Allocating cognitive resources

Figure 6-1 shows how resource allocation may be incorporated into the standard perception-cognition-action loop used to describe human decisions. Motor actions have traditionally been analyzed within the framework of Bayesian Decision theory (Kording & Wolpert, 2006; Maloney et al., 2007), but recently it has also been used for active sensing in vision. Najemnik and Geisler (2005) demonstrated that people strategically take into account the resolution of their visual field when they move their eyes in a visual search task. Here we argue that internal actions – the allocation of cognitive resources such as the allocation of visual attention, memory, and processing – should also be analyzed in terms of their effect on task performance and gain maximization.

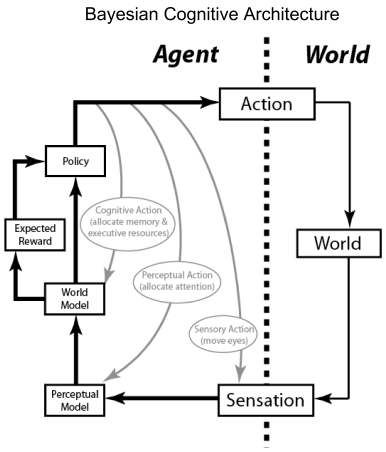

Figure 6-1: Bayesian decision theory applied to human cognition typically describes the cycle of sensation to perception and Bayesian estimation of the state of the world, through to policies for maximizing expected rewards and actions that affect the world. Our proposal is that in addition to to this action loop, there are similar processes for planning and executing "cognitive actions". For instance, saccades are a sensory action that determine what visual information will be received; visual attention can be considered a perceptual action that further determines how sensory information is processed. Memory is a strategic cognitive action that determines what information will be available in the future.

6.2 Memory allocation in multiple object tracking

Memory is useful insofar as it improves future actions with information from the past, so we would expect it to operate strategically in the context of a specific task. Vul et al. (2010) we explicitly modeled the strategic allocation of short term memory in a multiple object tracking task and compared it to human performance. At the computational level, when tracking objects we must determine which noisy observations at any given point in time should be associated with previously recorded object locations and trajectories. Although the computational-level description of this problem is sufcient to account for many commonly observed phenomena in human object tracking (effects of speed, spacing, predictability), it does not predict the characteristic tradeoff between the speed and the number of objects that humans can track. In this task, the role of memory is to maintain a record of previous object locations and trajectories, but all of this information cannot be recorded with innite precision; therefore, some objects should be remembered more precisely than others, depending on predicted future task demands. We show that an optimal, strategic allocation of a limited memory resource can account for the speed-number tradeoff in multiple object tracking. Furthermore, this model predicts human performance in additional experiments manipulating which objects should be remembered with more, or less, detail.

6.3 Strategic adjustment of sampling precision

Just as memory can be treated as strategic adjustment of the precision of state estimates to maximize future performance and gains, sample-based inference in general can be treated in this light. If the precision of sample-based decisions is adjusted to maximize expected gains, we would expect people to use more samples for decision that have higher stakes — do people make these predicted, optimal adjustments?

We can test this prediction within the literature on "probability matching" (Herrnstein, 1961; Vulkan, 2000). In these tasks, subjects choose alternatives with a frequency proportional to the probability of reward under that alternative. On our account, such "probability matching" arises from decisions made based on one sample – decisions based on more samples would correspond to Luce choice exponents greater than 1 (Luce, 1959). We can ask whether people adjust the number of samples they use as the stakes of a decision change? We tested the effect of higher stakes on the apparent number of samples used to make a decision in a more graded fashion within the set of experimental findings reviewed by Vulkan (2000). Specifically, we computed the average stakes of the decisions and an estimate of the number of samples subjects used to make those decisions (the Luce choice exponent) for each of the studies reviewed in Vulkan (2000). Our prediction is that when the stakes are higher (that is, when the difference in expected rewards between the maximizing and probability matching response strategies is large) subjects would use more samples for each decision, and thus would show a higher Luce-choice exponent.

Vul, Goodman, Griffiths, and Tenenbaum (submitted) found a significant positive correlation between the stakes of a decision and the Luce choice exponent of probability-matching behavior in those decisions. Thus, despite all of the other variation across studies and labs, when stakes are higher, people are closer to maximizing — they seem to use more samples per decision when it matters more.

6.4 Conclusion

Although the work presented in this discussion is preliminary, we think it makes important first steps towards a complete Bayesian cognitive architecture, based on two pillars. First, using sampling algorithms as process models for carrying out approximate Bayesian inference despite limited cognitive resources. Second, considering the allocation of cognitive resources as another problem for Bayesian decision theory, indicates how resource limitations may be treated within a rational analysis framework. There is much to be fleshed out along both research trajectories, however, these two pillars hold much promise for the foundations of a Bayesian cognitive architecture.

THIS PAGE INTENTIONALLY LEFT BLANK

References

Alais, B., & Blake, R. (2005). *Binocular rivalry*. Cambridge: MIT Press.

Alvarez, G., & Oliva, A. (2008). The representation of simple ensemble features outside the focus of attention. *Psychological Science*, *19*(4), 392-398.

Alvarez, G., & Oliva, A. (2009). Spatial ensemble statistis are efficient codes that can be represented with reduced attention. *Proceedings of the National Academic of Sciences*, *106*(73), 7345-7350.

Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Anderson, J., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*.

Ariely, D., Au, W., Bender, R., Budescu, D., Dietx, C., & Gu, H. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 13-147.

Ashby, F., Prinzmetal, W., Ivry, R., & Maddox, W. (1996). A foral theory of feature binding in object perception. *Psychological Review*, *103*(1), 165-192.

Averbach, E., & Coriell, A. (1961). Short term memory in vision. *Bell Systems Technical Journal*, *40*(309-328).

Baars, B. (1997). Some essential differences between consciousness and attention, perception and working memory. *Consciousness and cognition*, *6*(2-3), 363-371.

Battaglia, P., & Schrater, P. (2007). Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. *Journal of Neuroscience*, *27*, 6984-6994.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behavior. *Trends in Cognitive Sciences*, *11*, 118-125.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*, 700-765.

Botella, J., Arend, I., & Suero, M. (2004). Illusory conjunctions in the time domain and the resulting time-course of the attentional blink. *Spanish Journal of Psychology*, *7*(1), 63-68.

Botella, J., & Eriksen, C. (1992). Filtering versus parallel processing in rsvp tasks. *Perception and psychophysics*, *51*(4), 334-343.

Botella, J., Garcia, M., & Barriopedro, M. (1992). Itrusion patterns in rapid serial visual presentation tasks with two response dimensions. *Perception and psychophysics*, *52*(5), 547-552.

Botella, J., Suero, M., & Barriopedro, M. (2004). A model of the formation of illusory conjunctions in the time domain. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(6), 1452-1467.

Brainard, D. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433-436.

Brainard, D., & Freeman, W. (1997). Bayesian color constancy. *Journal of the Optical Society of America*, *14*(7), 1393-1411.

Broadbent, D. (1958). *Perception and communication.* London: Pergamon Press.

Brown, S., & Steyvers, M. (2008). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Bruner, J., Goodnow, J., & Austin, G. (1956). *A study of thinking.* New York, NY: Wiley.

Bundasen, C., Kyllingsbaek, S., & Larsen, A. (2003). Independent encoding of colors and shapes from two stimuli. *Psychonomic Bulletin and Review*, *10*(2), 474-479.

Carrasco, M. (2006). Covert attention increases contrast sensitivity: Psychophysical, neurophysiological and neuroimaging studies. *Progress in Brain Research*, *154*, 33-70.

Cepeda, N., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distribution of practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354-380.

Cepeda, N., Vul, E., Rohrer, D., Wixted, J., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095-1102.

Charniak, E. (1995). *Parsing with context-free grammars and word statistics* (Tech. Rep.). Providence, RI: Brown university.

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for bayesian cognitive science.* UK: Oxford University Press.

Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287-291.

Chong, S., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception and psychophysics*, *67*(1), 1-13.

Chun, M. (1994). Processing deficits in rsvp: the attentional blink and repetition blindness. *dspace.mit.edu*.

Chun, M. (1997). Temporal binding errors are redistributed by the attentional blink. *Perception and psychophysics*, *59*(8), 1191-1199.

Chun, M., & Potter, M. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 109-127.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559-583.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. In *Advances in neural information processing systems* (Vol. 21). Vancouver, Canada.

Deese, J., & Kaufman, R. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, *54*(3), 180-187.

Doucet, A., De Freitas, N., & Gordon, N. (2001). *Sequential monte carlo methods*. New Yok: Springer.

Estes, W. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134-140.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational world learning. *Psychological Science*.

Galton, F. (1889). *Natural inheritance*. London: MacMillan.

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450-451.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Gershman, S., Vul, E., & Tenenbaum, J. (2010). Perceptual multistability as markov chain monte carlo inference. *Advances in Neural Information Processing Systems*.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press, USA.

Gold, J., & Shadlen, M. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*, 390-394.

Gold, J., & Shadlen, M. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*, 299-308.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

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

Griffiths, T., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354 – 384.

Griffiths, T., & Tenenbaum, J. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767-773.

He, S., Cavanagh, P., & Intrilligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, *383*(6598), 334-337.

Herrnstein, R. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267.

Hertwig, R., & Pleskac, T. (2010). Decisions from experience: Why small samples? *Cognition*, *115*, 225-237.

Hirt, E., & Markman, K. (1995). Multiple explanation: a consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*(6), 1069-1086.

Huang, L., & Pashler, H. (2007). A boolean map theory of visual attention. *Psychological Review*, *114*(3), 599-631.

Huang, L., Treisman, A., & Pashler, H. (2007). Characterizing the limits of human visual awareness. *Science*, *317*(5839), 823-825.

Humphrey, G., & Goodale, M. (1998). Probing unconscious visual processing with the mccollough effect. *Consciousness and cognition*, *7*, 494-519.

Huttenlocher, J., Hedges, L., & Vevea, J. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*, 220-241.

Isenberg, L., Nissen, M., & Marchak, L. (1990). Attentional processing and the independence of color and orientation. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 869-878.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295-1306.

James, W. (1890). *The principles of psychology*. Dover publications.

Kahneman, D., Treisman, A., & Gibbs, B. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175-219.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinions in Neurobiology*, *13*(2), 150-158.

Kikuchi, T. (1996). Detection of kanji words in a rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(2), 332-341.

Kinchla, R., & Smyzer, F. (1967). A diffusion model fo perceptual memory. *Perception and psychophysics*, *2*(219-229).

Knill, D., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Koch, C., & Tsychiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in Cognitive Sciences*, *11*(1), 16-22.

Kording, K. (2007). Decision theory: What should the nervous system do? *Science*, *318*(5850), 606.

Kording, K., & Wolpert, D. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Lamme, V. (2003). Why visual attentiona nd awareness are difference. *Trends in Cognitive Sciences*, *7*(1), 12-18.

Landman, R., Spekreijse, H., & Lamme, V. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149-164.

Laplace, P. (1818). *Theorie analytique des probabilites*. Paris, France: Impreimerie Royale.

Levy, R., Reali, F., & Griffiths, T. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in neural information processing systems* (Vol. 21). Whistler, Canada.

Liberman, A., KS, H., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358-368.

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movememnt and depth: anatomy, physiology and perception. *Science*, *240*, 740-749.

Luce, R. (1959). *Individual choice behavior*. New York, NY: Wiley.

Ma, W., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432-1438.

Maloney, L. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfield (Eds.), *Perception and the physical world* (pp. 145–189). New York, NY: Wiley.

Maloney, L., Trommershauser, J., & Landy, M. (2007). Questions without words: A comparison between decision making under risk and movement planning under risk. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 297–313). New York, NY: Oxford University Press.

Marr, D. (1982). *Vision*. Cambridge: MIT Press.

McKenzie, C. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and bayesian inference. *Cognitive Psychology*, *26*, 209–239. Available from `http://psy.ucsd.edu/ mckenzie/McKenzie1994CogPsych.pdf`

Medin, D., & Shaffer, M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Monheit, M., & Johnson, J. (1994). Spatial attention to arrays of multidimensional objects. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 691-708.

Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*, 1133-1147.

Mozer, M., Shettel, M., & Vecera, S. (2005). Top-down control of visual attention: A rational account. *Advances in Neural Information Processing Systems*.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable achoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142.

Najemnik, J., & Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387-391.

Necker, L. (1832). Observations on some remarkable phenomenon which occurs in viewing a figure of a crystal or geometrical solid. *Phil. Mag. J. Sci*, *3*, 329-337.

Nieuwestein, M., Chun, M., & Lubbe, R. van der. (2005). Delayed attentional engagement in the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1463-1475.

Nieuwestein, M., & Potter, M. (2006). Temporal limits of selection and memory encoding: A comparison of whole versus parial report in rapid serial visual presentation. *Psychological Science*, *17*(6), 471-475.

Nissen, M. (1985). Accessing features and objects: Is location special? *Attention and Performance XI*.

Nosofsky, R., Palmeri, T., & McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53-79.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central

bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 358-377.

Pelli, D., Palomares, M., & Majaj, N. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, *4*(12), 1136-1169.

Posner, M. (1994). Attention: the mechanisms of consciousness. *Proceedings of the National Academic of Sciences*, *91*(16), 7398-7403.

Posner, M., Snyder, C., & Davidson, B. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160-174.

Prinzmetal, W., Henderson, D., & Ivry, R. (1995). Loosening the constraints on illusory conjunctions: assessing the roles of exposure duration and attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1362-1375.

Prinzmetal, W., Ivry, R., Beck, D., & Shimizu, N. (2002). A measurement theory of illusory conjunctions. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 251-269.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Raymond, J., Shapiro, K., & Arnell, K. (1992). Temporary suppression of visual processing in an rsvp task: an attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849-860.

Reeves, A., & Sperling, G. (1986). Attenton gating in short-term visual memory. *Psychological Review*, *93*(2), 180-206.

Ripley, B. (1987). *Stochastic simulation*. New York: Wiley.

Robert, C., & Casella, G. (2004). *Monte carlo statistical methods*. New York: Springer.

Robertson, L. (2003). Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, *4*, 93-102.

Sanborn, A., & Griffiths, T. (2008). Markov Chain Monte Carlo with People. In *Advances in neural information processing systems* (Vol. 20). MIT Press.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 726–731). Vancouver, Canada.

Schneider, A., Oppenheimer, D., & Detre, G. (2007). Application of Voting Geometry to Multialternative Choice. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 635–640). Nashville, TN.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, *15*(11), 720-728.

Shafto, P., Kemp, C., Bonawitz, E., Coley, J., & Tenenbaum, J. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*(2), 175-192.

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

Shepard, R. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703.

Shi, L., Feldman, N., & Griffiths, T. (2008). Performing bayesian inference with exemplar models. In *Proceedings of the 30th annual conference of the cognitive science society.* Washington DC, USA.

Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. (in pressa). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin and Review*.

Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. (in pressb). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin and Review*, *404*.

Shih, S., & Sperling, G. (2002). Measuring and modeling the trajectory of visual spatial attention. *Psychological Review*, *109*(2), 260-305.

Sincich, L., & Horton, J. (2005). The circuitry of v1 and v2: integration fo color, form and motion. *Annual Reveiw of Neuroscience*, *28*, 303-326.

Sobel, D., Tenenbaum, J., & Gopnik, A. (2004). Chidlren's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333.

Srinivasan, R. (2002). *Importance sampling - applications in communications and detection.* Berlin: Springer-Verlag.

Stewart, N., Chater, N., & Brown, G. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Stewart, T. (1999). Principles of forecasting: A handbook for researchers and practitioners. In J. Armstrong (Ed.), (chap. Improving reliability of judgmental forecasts). Kluwer Academic Publishers.

Steyvers, M., Griffiths, T., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*(7), 327-334.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Cambridge: Harvard University Press.

Strasburger, H. (2005). Unfocused spatial attention underlies the crowding effect in indirect form vision. *Journal of Vision*, *5*(11), 1024-1037.

Sundareswara, R., & Schrater, P. (2007). Dynamics of perceptual bistability. *Journal of Vision*, *8*(5).

Suroweicki, J. (2004). *The wisdom of crowds.* New York, NY: Random House.

Swets, J., Tanner, W., & Birdsall, T. (1961). Decision processes in perception. *Psychological Review*, *68*, 301-340.

Taleb, N. (2008). *The black swan.* New York, NY: Random House.

Tenenbaum, J. (1999). Bayesian modeling of human concept learning. *Advances in Neural Information Processing Systems*.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, *14*(4), 411-443.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*(1), 107-141.

Trommershauser, J., Maloney, L., & Landy, M. (2003). Statistical decision theory and rapid, goal-directed movements. *Journal of the Optical Society A*, 1419-1433.

Tsal, Y. (1989). Do illusory conjunctions support the feature integration theory? a critical review of theory and findings. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 394-400.

Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Van Rullen, R., & Koch, C. (2003). Is perception discrete or continuous. *Trends in Cognitive Sciences*, *7*(5), 207-213.

Vul, E., Frank, M., Alvarez, G., & Tenenbaum, J. (2010). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems*.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (n.d.). One and done? optimal decisions from very few samples. *Proceedings fo the 31st Annual Meeting of the Cognitive Science Society*.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (submitted). One and done? optimal decisions from very few samples.

Vul, E., Hanus, D., & Kanwisher, N. (2008). Delay of selective attention during the attentional blink. *Vision Research*, *48*(18), 1902-1909.

Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: Selection is probabilistic; responses are all or none samples. *Journal of Experimental Psychology: General*, *4*, 546-560.

Vul, E., & MacLeod, D. (2006). Contingent aftereffects distinguish conscious and preconscious color processing. *Nature Neuroscience*, *9*(7), 873-874.

Vul, E., Nieuwestein, M., & Kanwisher, N. (2008). Temporal selection is suppressed, delayed, and diffused during the attentional blink. *Psychological Science*, *19*(1), 55-61.

Vul, E., & Pashler, H. (2007). Incubation is helpful only when people are misled. *Memory and Cognition*, *35*, 701-710.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Vul, E., & Rich, A. (in press). Independent sampling of features enables conscious perception of bound objects. *Psychological Science*.

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

Wallsten, T., Budescu, D., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*(3), 243-268.

Weichselgartner, E., & Sperling, G. (1987). Dynamics of automatic and controlled visual attention. *Science*, *238*(4828), 778.

Welford, A. (1952). The 'psychological refractory period' and the timing of high-speed performance – a review and a theory. *British Journal of Psychology*, *43*, 2-19.

Whitely, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty is used to guide statistically optimal decisions. *Journal of Vision*, *8*(1-15).

Wilson, H., Blake, R., & Lee, S. (2001). Dynamics of travelling waves in visual perception. *Nature*, *412*, 907-910.

Wixted, J., & Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409-415.

Wolfe, J., & Cave, K. (1999). The psychophysical evidence for a binding problem in human vision. *Neuron*, *24*, 11-17.

Xu, F., & Tenenbaum, J. (2007). Sensitivity to sampling in bayesian word learning. *Developmental Science*, *10*(3), 288–297.

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation and attention. *Neuron*, *46*, 681-692.

Yuille, A., & Bülthoff, H. (1996). Bayesian decision theory and psychophysics. In D. Knill & W. RIchards (Eds.), *Perception as bayesian inference* (pp. 123–161). Cambridge, MA.