Perception of Probabilities which are Subject to Change

by

Julia Schirmeister

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Arts
in
Psychology

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners

I understand that my thesis may be made electronically available to the public.

Abstract

To navigate stochastic and changing environments, people need to keep track of ongoing probabilities as those probabilities are subject to change. Two distinct theories of mental-model updating are compared. In trial-by-trial updating models, every sample is immediately integrated into a working estimate of the probability. In change-point detection, a single estimate of the probability is maintained until evidence accumulates to reject that model to adapt a new model. Disentangling these theories of updating frequencies has been difficult due to a confound found in previous tasks. Participants have been given their last response as their default response, and this has made it easier for them to maintain the same estimate rather than update it. This favours change-point models. To address whether response-maintenance is due to the extra effort it takes to update a response, participants were separated into two groups. In the Automatic condition, participants were given their old response as default. In the Manual condition, participants were given no default and were asked to generate a new estimate of the probability every trial. While offering a default response was found to partially explain response maintenance in previous tasks, it did not fully explain it. Participants in the Manual group showed spontaneous meticulous response maintenance over long series of trials despite being asked to respond anew every trial. This suggests that the hypothesis-testing strategy developed in the change-point detection literature is a fundamental component of probability estimation and is not an artifact of previous task designs.

Acknowledgements

I would firstly like to offer the most deep and sincere thanks to my supervisor Britt Anderson who got me started on this project and offered me so much guidance, support, and insight over the course of my thesis.

I would also like to thank my two readers, each of whom taught courses I treasured very much and from which I gained a lot of insight.

I would also like to thank my lab members who have been very helpful and supportive; they have offered really useful feedback and given me a lot of new ideas to work with.

Lastly, I would also like to extend my thanks to my family and my friends who have offered me encouragement and support.

Table of Contents

List of Figures vii

1 Introduction 1

1.1 Historical Precedent: Delta-Rule . . . 3

1.2 Change-point Detection . . . 5

1.2.1 IIAB Model . . . 8

1.3 Model Comparison . . . 10

1.4 Current Study . . . 11

2 Methods 14

2.1 Demographics . . . 14

2.2 Preliminaries and Practice Trials. . . . 14

2.3 Experimental Procedures . . . 15

2.4 Task Details . . . 15

2.4.1 Practice Trials . . . 16

2.4.2 Experiment Trials and Conditions . . . 16

2.4.3 Exclusion . . . 17

3 Results 19

3.1 Survey Responses . . . 19

3.2 Practice Trials . . . 20

3.3 Effects of Condition on Task Performance 20
3.4 Individual Variability within Conditions 22

4 Discussion **28**
4.1 Study Overview 28
4.2 Explaining Spontaneous Response Maintenance 31
4.3 Relating Response Maintenance to Hypothesis Testing 32
4.4 Expanding the Notion of Hypothesis Testing 34
4.4.1 Incremental Single Model Adjustment versus Model Comparison . . 35
4.5 Conclusion 42

References **43**

List of Figures

1.1 Illustration of a step-hold pattern in participant estimates of probability. . 6

2.1 Experimental display 15
2.2 Histograms of participant accuracy separated by block. 18

3.1 Boxplot showing group differences in total number of adjustments 21
3.2 Compared model fits for Manual participant added adjustment distributions 23
3.3 Performance over task of Manual participants who made the fewest adjustments 24
3.4 Performance over task of Automatic participants who made the most adjustments 25
3.5 Number of sample-consistent adjustments by number of total adjustments . 26
3.6 Proportion of sample-consistent adjustments by number of total adjustments 27

Chapter 1

Introduction

The ability to learn the statistics of environments is crucial for survival. While it may seem unassuming at first, the importance of the ability to represent relative probabilities of events is quite profound. Expectation can resolve ambiguities, speed processing, or be used to make predictions about future events. Even basic perceptual processes use learned probability distributions. Something as simple as resolving a signal to a particular event uses probabilistic information. As an illustrative example, take the ability to parse speech signal into phonetic segments. Any given language has a set of segments (consonants and vowels) which are units combined to form words. When language-users are given acoustic signal, the sound they report hearing is pulled towards the centers of phonemic activity [1]. This is an advantage if an utterance is noisy, because the most likely speech given the signal will be what is heard [2].

Understanding how probability distributions are learned using limited sets of samples is an open question. It is further complicated by the fact that probabilities change; they change over times, places, and abstract spaces. In order for humans to adapt flexibly to different environments, they need to be able to respond appropriately to these conditional probabilities. Returning to the example of speech, the same phonemic signal will be processed differently under different contexts. Phonological explanations for speech productions are primed by higher order information such as semantic context or the speech patterns of the particular speaker [3].

Learning whether conditions have changed and new probabilities apply is not an easy problem. Changes in the hidden process generating samples may have no outward indication. When this state information is absent, learning different probabilities is constrained to detecting differences in the stochastic hidden process. That is, in distinguishing between

unlikely episodes of events, and changes in probabilities. Changes in probabilities can occur at any level of stochasticity. A first order stochastic process is a process where there is a single fixed probability distribution on possible events generating samples. A second order stochastic process is a process where this probability distribution itself varies. Any higher level of stochasticity is a change in the super-structure that determines the changes in levels of stochasticity below.

In a laboratory setting, people are quite adept at rapidly responding to hidden changes in at least doubly stochastic processes [4, 5, 6, 7, 8, 9, 10, 11]. This ability is found in mice as well as humans [12, 13]. This cross-species ability suggests that detecting change in probabilities is a fundamental perceptual skill. In these nonstationary probability estimation experiments, participants are given a series of samples drawn from a hidden generative process. This process is subject to hidden changes at distinct change-points. At a change-point, the active probability switches to a new distribution and samples start to be generated from the new distribution. Participants respond readily to these sudden changes even if they are unmarked by other events.

Given nonstationary environments where probabilities change, two hypotheses are contrasted to explain how people update their estimates of probabilities. Either people use every sample to update their estimate of the probability, or they fix an estimate and accumulate samples which determine whether the probability has since changed. This distinction corresponds loosely to the difference between two learning principles identified by Gallistel et al. [4]. These are referred to as trial-by-trial updating (where one trial corresponds to one sample and updating occurs every trial) and change-point detection (where probabilities are taken to be static over many trials (runs) until change-points mark discrete jumps from one probability distribution to another).

Until recent work by Gallistel et al. [4], frequency of mental model updating has not been specifically investigated. Instead, for reasons discussed below, trial-by-trial updating has been assumed. As identified by Gallistel et al., it is quite difficult to disentangle the two hypotheses given existing literature. One reason is that, in early versions of this task, participants were asked to guess the value of the next sample. This is instead of more directly being asked to describe the distributions directly (e.g. [10, 8, 14]). This required additional work by the researcher to infer the participant estimates of the probability distribution. This additional work implies a number of assumptions needing to be made by the researcher. These assumptions may not have an established resolution. For example, it is still unknown how guesses generated from estimated probability distributions relate to those distributions [15].

Gallistel et al. argue that asking for guesses particularly impedes inference to probabil-

ity estimation in the most common nonstationary probability estimation task. The simplest stochastic process, the Bernoulli process, is often used in these probability estimation experiments. The Bernoulli process is defined by a single parameter, p, which determines the probabiliy of one of two binary outcomes represented by the set [0,1]. Given only binary-state estimates of the next dot, to infer a probability estimate, researchers need to integrate over multiple responses to determine the average estimated binary sample. This has poor temporal resolution regarding the frequency with which participants update their estimates of the probabilities.

Apart from a single early case [11], only a recent emerging literature, following Gallistel et al. [4], asks participants to estimate the Bernoulli process parameter with each new sample. This provides the temporal resolution required to distinguish the two hypotheses proposed above. In these tasks, participants are asked to imagine they are drawing coloured balls out of an urn. They are asked to estimate the proportion of coloured dots in the urn determining their chance of drawing either colour. To estimate the proportion, they are asked to set a slider at the bottom of the display. Each end of the slider represents 100% of the dots being of either colour and any setting in-between those extremes represents the estimated proportion of coloured dots in the urn. This way, with each new sample, participants can adjust their estimates of the probability over a continuous scale.

This new task paradigm has allowed for better distinction between mental model updating which occurs with each sample and mental model updating which is delayed until the previous mental model is found inadequate. To understand the strengths of the competing hypotheses, the next section explores the history underlying the early acceptance of trial-by-trial models and the emergence of a competing change-point detection model.

1.1 Historical Precedent: Delta-Rule

In the past, probability estimates of stepwise nonstationary stochastic processes have been explained by use of the delta-updating rule to generate estimates of the probability [10, 16, 8, 9, 14]. The delta-updating model is one of a set of models called trial-by-trial updating models. On every trial where a new sample is generated or an event occurs, the running estimate kept by the model is updated to make the event more likely.

$$p_{t+1} = p_t + \alpha\delta$$

According to the delta rule, probability p_{t+1} at time t is the previous estimate of the probability p_t plus α the learning rate times the δ error.

In the delta model, the error of the model given the most recent sample is used to update the model. The previous model will have an associated expectation which the actual outcome, the particular sample, will be some distance from. The difference between the expected and actual outcome will determine the fastest way which the probability estimate needs to be improved to make the most recently observed data more likely. This is the error. It is multiplied by a learning rate and added to the previous estimate of the probability. This produces the new estimate. Delta-model learning can be understood as a stochastic approximation of gradient ascent [17].

An advantage of the delta rule model is its computational simplicity. This makes it easy to extend to possible neurological models which could realize its principles of operation. Neurological models exist which connect the delta-updating rule to prediction errors registered in the anterior cingulate cortex [18]. Another advantage of the delta rule model which makes it neurologically plausible is that it also does not require extensive memory to store recent events. All retained information from previous samples is contained recursively in the current estimate of the probability.

A disadvantage of the delta-rule updating model is that it introduces a trade-off between rapid response to changes in probability and steady accumulation of more information over time [4]. With a fixed learning rate, the weight of any particular sample used in the current estimate of the probability decreases geometrically with distance from the most recent sample. This fixes how much information is used for an estimate at any given time. If the learning rate is low, the delta model accumulates a lot of information to generate its current estimate of the probability. If the learning rate is high, the delta model responds rapidly to unusual events which may mark discrete jumps in the underlying generative process.

This trade-off can be mitigated, however, if the learning rate is not fixed. Nassar et al. [8] for example, simplify optimal Bayesian learning and produce a model with a variable learning rate. The learning rate is a function of other estimated variables which determine how confident one can be about the current estimate of the probability. The more confidence in a current estimate, the less any new sample sways this estimate. In the model proposed by Nassar et al., the learning rate is a function of the current estimated run-length (a longer run means a lower learning rate) and the estimated hazard rate or expected frequency of change-points (a higher hazard rate means a higher learning rate). Thus, because these estimates are also updated trial-by-trial, the learning rate varies along with estimates of the probability.

Human estimates of probability indeed show variable learning rates. Moreover, these learning rates vary with the learning rates of the optimal Bayesian observer. The more

narrow the posterior distribution around the optimal Bayesian estimate, that is, the more confident the optimal Bayesian learner is about their particular estimate of the probability, the lower a human learning rate will be. Because the model provided by Nassar et. al is a reduced optimal Bayesian change-point detection algorithm, human learning rates are also predicted by this adapted delta model [8].

Another strategy for adaptive learning rates is a two-kernel model [4]. Instead of having a single estimate of the probability as determined by a delta-rule model with a fixed learning rate, there are two ongoing estimates both using the delta-rule model but each using a different learning rate. The slow learning rate model is used by default. This is because, most of the time, it is desirable to use as much information as possible to determine one's current estimate of the probability. Whether to switch from one model to the other is determined by the magnitude of the difference between the two learning models. A large difference between the two estimates signals a steep change in the fast-learning model. This is likely to have been caused by the probability having changed.

1.2 Change-point Detection

Despite being computationally simple and neurologically realizable, delta-updating trial-by-trial models have been met more recently with criticism. In particular, Gallistel et al. [4] offers a series of strong arguments against them. These arguments are made on both theoretical and empirical grounds. On theoretical grounds, Gallistel et al. identify a number of shortcomings trial-by-trial models have which would make their use disadvantageous to an adaptive organism. On empirical grounds, Gallistel et al. find distinct qualitative patterns in human estimates of probability which they could not reproduce given the delta-updating model.

One of the most important of these empirical phenomena identified by Gallistel et al. is the low frequency with which humans update their estimates of probability. Participants adjust their estimates less often than on every trial. Instead of changing responses with each new piece of evidence, participants hold the same estimate of the probability steady over longer series of trials. Only intermittently, after runs of a maintained estimate, do participants then use large steps to update their estimates. Gallistel et al. identify this as a step-hold pattern which is illustrated in Figure 1.1. This step-hold response pattern is a robust finding across previous research [11, 6, 5]. Estimates of the probability are maintained more often than the Bayesian benchmark [6].

The distinct step-hold pattern of participant estimates, Gallistel et al. [4] argues is

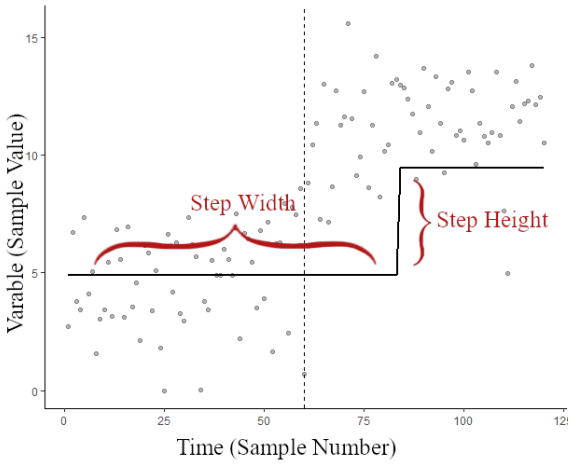

Figure 1.1: Illustration of a step-hold pattern in participant estimates of probability. In this representative task, a participant is asked to estimate the mean value of a Gaussian distribution generating the samples they observe. On the y-axis is the value of the sample observed. On the x-axis is trial number. Each dot represents one sample drawn from the Gaussian distribution on one trial. The solid horizontal lines represents the participant's estimate of the mean of the Gaussian distribution on each trial. At the half-way point, indicated by the vertical dashed line, the mean of the true Gaussian changes. Only after several trials have passed does the participant's estimate of the mean also change. This is because they were accumulating evidence over many samples to conclude that the mean has changed. The long period of trials before a change is made is the step-width. The change in estimate is the step-height.

evidence of discrete hypothesis testing. According to a hypothesis testing account of probability estimation, people adopt mental models as static estimates of the state of the world which explains current observations. These estimates of the underlying generative distribution remain fixed over samples. It is not until enough evidence accumulates that would be inconsistent with a current working model that a person finally rejects the previous distinct mental model and adapts a new model.

The principle of hypothesis testing can be related to temporally extended doubly stochastic processes via change-point detection algorithms [4]. In change-point detection for time-series, the discrete hypotheses are the probabilities active for discrete runs between distinct change-points. For the active estimate of the running probability, one hypothesis is adapted to explain the set of samples seen since the most recent purported change-point. This hypothesis is contrasted with the probability that a change-point has occurred and the probability has changed since. The defining feature of change-point models is that they insert discrete change-points between hypothesized runs of fixed probability.

Step-hold response patterns are taken as evidence for the suspending of mental model updating until sufficient evidence has accumulated to reject the working hypothesis. This suspension of belief-updating recalls the distinction between surprise and updating [19]. Surprise relates to the unlikeliness of an event given a working probability model. O'Reilley et al. characterize it by an event's Shannon information. Surprise does not need to lead to any updating. Updating is the difference between the prior estimate of the probability before the recent sample and posterior estimate of the probability after the new information alters the estimate. It is thus related to the impact the observation had on the running estimate of the probability. O'Reilley et al. measure it using the Kullback-Liebler (KL) divergence between the previous and new estimate.

O'Reilley et al. [19] used a saccade task to dissociate two pathways for surprise and updating. Surprise was localized to the Inferior Parietal Lobule (IPL), corresponding to immediate preparation for a change in motor response to unexpected stimuli on the current trial. In contrast, updating events involved the Intraparietal Sulcus (IPS) and Anterior Cingulate Cortex (ACC). These brain regions were theoretically recruited for more permanent alteration of expectation in future trials. This learning was theoretically being mediated through the noradrenergic system from the locus coeruleus. The identification of two separate pathways for surprise and updating supports the theory that mental model updating can be suspended if the running hypothesis has not been rejected.

Gallistel et al. [4] argue that there are other empirical phenomena which can only be explained by change-point detection over trial-by-trial updating. Humans can have second thoughts and take back an inserted change-point from the time series; they can change

their mind that the probability has changed and revert back to their old estimate of the probability. In the task given by Gallistel et al., participants were asked to press a button if they detected the probability had changed. The first five participants all spontaneously reported at the conclusion of the task that there were times they changed their mind about pressing the button. While they believed for some trials that the probability had changed, retrospectively they inferred that they had only witnessed an unlikely series of events and they should have maintained their old estimate. This is only consistent with a model that proposes distinct change-point detection.

Second-thoughts reveal a deeper phenomenon active in participant estimates of probability over time: retrospective inference. Gallistel et al. [4] argue that change-point detection is particularly good for information compression and retrospective inference. Using the delta rule, information over time is stored as a series of estimates that are contingent on the value of the data on any given trial. Memory for this series of estimates is unique to each event. In contrast, when discrete change-points are inserted throughout the time series, memory storage is a much smaller set of summary statistics representing blocks of time. As Gallistel et al. [4] summarizes, the efficiency of change-point detection is similar to a principle used for lossless data compression. Redundant signal can be removed from data if the common information is stored once. Instead of maintaining a series of similar estimates across a time series, change-point detection stores summary statistics and the differences between runs. This way, recollection of the series of events, is compressed to a shorter list of summary statistics.

Apart from the empirical phenomena, a further argument for change-point detection is that it is a strategy that would be an advantage in a natural setting and thus it had a potential evolutionary function. Similar to a variable learning rate, change-point detection is sensitive to the amount of data that should be considered in the current estimate of the probability. However, unlike a variable learning rate, it considers precisely only those data points which are relevant to the current estimate of the probability. In contrast, for example, in the model proposed by Nassar, the current estimate of the probability is approximated to be a weighted estimate of the marginal of all possible run-lengths.

1.2.1 IIAB Model

This line of reasoning clearly establishes a strong argument that humans use a version of change-point detection to generate their estimates of probabilities. However, change-point detection algorithms can take on many forms. An online ideal Bayesian observer for optimal change-point detection on time-series data has been developed [20, 21]. However,

as the time series expands, the algorithm rapidly grows in computational complexity and memory demand. Given a known hazard rate, computations increase linearly with run-length [20]. However, if the hazard rate needs to be estimated, computational complexity increases exponentially [21]. This rapid growth in computational expense arguably makes it neurologically implausible and thus it is unlikely to be the strategy adopted by humans.

To vastly reduce this complexity of the optimal (full Bayesian) solution, Gallistel et al. [4] offers a computationally elegant approximation of optimal Bayesian change-point detection. This model is referred to as the 'If It Aint Broke' (IIAB) model. In the IIAB model, estimates of the probability are not changed unless the current estimate is broke, or in other words, not sufficiently predictive of the most recent data. This title distinguishes it from a trial-by-trial updating model where updating is constant. In the IIAB model, if the current estimate is found to be broke, then a change-point is inserted into the time series. Around each change-point, the probabilities estimates to be active during those runs are the optimal estimates given those run-lengths.

There are key differences between optimal Bayesian change-point detection and the IIAB model. Foremostly, in optimal Bayesian online change-point detection, when new change-points are inserted, this takes into account every possible intersection of all previous possible change-point locations. Any new change-point is the change-point that makes the previously seen data most likely given all possible previous change-point locations. Online consideration of all possible previous runs is made possible by recursive storage of information in a message-passing updating algorithm [20].

In contrast, in the IIAB model, there is no message-passing algorithm; there are no nodes representing all possible run-lengths for the current trial. Instead, a single estimate of the progression of change-points over the task is considered. On any given trial, if a decision is made about inserting a change-point, only three hypotheses are compared. These possible changes include whether to add a change-point since the most recent change-point, whether to take away the last change-point, or whether to move the most recent change-point somewhere else.

Further detail elucidates the full algorithm. Online IIAB inference is a two-stage process. The first stage is a threshold process, which prevents probability re-estimation and adjustment from occurring every trial. The second stage is the change-point estimation subroutine outlined above, which determines whether to insert, remove, or shift a change-point.

In the IIAB model, the threshold procedure first determines whether the model is broke and needs adjustment. A statistic, 'E', is calculated by taking the product of two variables. One portion of the product is the KL divergence (see [22]) between the estimated

probability distribution and the observed distribution of the most recent samples. The other part of the product is the estimated run-length or in other words the number of trials which have occurred since the last purported change-point. This product is intended to represent the probability that there has been a new change-point since the previous one. If the product, E, exceeds a threshold, the algorithm initiates its updating subroutine. E is left as a free-parameter to be fit to participant data.

The subroutine also is a two-step process. It first tests whether to insert a change-point and then it tests whether to remove the last change-point. Both tests are implemented as Bayes Factors (see [23]). Each Bayes Factor compares a one-parameter model to a three-parameter model. The one parameter model represents the hypothesis that there has been no change-point. The one parameter to be fit is the estimate of the probability for the set of samples seen since the last change-point. The three-parameter model represents the hypothesis that there is one more change-point to be inserted. Two of the parameters are the probabilities surrounding the change-point and the remaining parameter is the location of the change-point. A threshold on the Bayes Factor to insert or remove a change-point is another free-parameter in the IIAB model.

1.3 Model Comparison

The IIAB model substantiates a learning algorithm, based on change-point detection, which could explain participant estimates of probabilities over time series. Being a fully developed learning algorithm with free-parameters, the estimates produced by the IIAB model can be compared with estimates produced by delta-updating models.

As reviewed, the standard delta-updating model was criticized for updating too frequently, and Gallistel et al [4] rule it out by this account. However, the general delta-rule learning algorithm can be appended to introduce hypothesis testing and the insertion of change-points. To append the model and prevent constant updating, a simple threshold procedure can be added to a delta-updating model, transforming it into a two-stage process. The first, new added step, determines whether to update the estimate of the probability to the estimate made by the delta-updating model. Unlike the IIAB model where estimates are optimal given change-point locations, in a two-step delta-model, estimates around inserted change-points are taken from the delta-rule updating model. Across several studies, adding such a threshold to a trial-by-trial updating model does improve the fit of its estimates [24, 6, 7]. This suggests that it is possible that human estimates are updated trial-by-trial but a threshold prevents those updates from becoming the current working estimate.

Gallistel et al. [4] argue that a simple threshold procedure is not sufficient to explain participant mental model updating. They argue that there are qualitative behavioral patterns in participant adjustments to their previous estimates that cannot be replicated using a two-step delta-updating model. Forsgren et al. counter that this argument was made too hastily. They argue that the model comparison conducted by Gallistel et al. [4] did not sufficiently explore the available parameter space for their model fits. Qualitative behavioral patterns can indeed be reproduced by trial-by-trial model updating algorithms given a more exhaustive exploration of the available parameter space.

To further bolster the argument for two-step trial-by-trial updating methods, Forsgren et al. argue that Gallistel neglected to quantitatively compare the fit of different models participant data. Comparing the IIAB model to the delta model with various added threshold procedures, Forsgren et al. found that by far the best fitting learning algorithm was in fact a delta-learning model with an added drift-diffusion threshold procedure.

Because the best fit to participant estimates was the delta-updating model, Forsgren et al. argue that trial-by-trial probability estimation has not been ruled out as the method by which people update their estimates of probabilities.

One argument is of particular interest to us, and is the focus of the present study. Forsgren et al. critique the use of the identified step-hold pattern as an argument for change-point detection. Instead of this being evidence of mental model maintenance, Forsgren et al. offer an alternative account. In previous task designs, on each trial as participants are asked to update their estimate of the probability, they are given their old response as their default response. This Forsgren et al. argue, introduces a potential confound. It is more effortful to update one's response on the slider than it is to leave it in its default position. This may introduce a motor effort-to-respond threshold which prevents participants from making adjustments to their probability estimates. This would produce the same step-hold patterns originally identified by Gallistel et al. Thus, participant response maintenance may not reflect true suspended belief updating, but rather may reflect a threshold introduced by the motor cost of adjusting the slider. This potential confound in previous task designs further obscures the original question of whether evidence is immediately integrated into working mental models or whether hypotheses are maintained over samples and discretely switched between.

1.4 Current Study

The current study was designed to address this confound. The same Bernoulli estimation task was adapted from previous research. Participants were split into two conditions. In the

first condition, as in previous research designs, participants were given their last response as their default response on every single trial. In the second condition, participants were given no default response at all; they were asked to respond anew every trial. The removal of the default response was to standardize the motor procedure required for both maintaining and updating a response. It also removed the implicit suggestion to adapt a previous estimate as one's current estimate.

Previous task designs have also been administered over ten blocks of 1000 trials. This results in participants eventually completing 10,000 trials. This also introduces the same potential issue that belief updating is confounded with the effort it takes to update one's response. Such a long and tedious set of sessions may decrease vigilance over the course of the full task. This may reduce slider adjustments, which Gallistel et al. indeed did note occurred over the course of the task although they explained this as learning the hazard rate. To mitigate this confound, participants in our study were only given 999 trials. In part, this decision was made because participants were asked to do the study at home alone, where they may have been distracted by other tasks during the study had it taken too long.

These changes to the task design were introduced in order to determine whether increased motor effort to update responses fully explains the response-maintenance step-hold pattern found robustly in previous versions of this task. If it is the case that response maintenance persists despite changes in the task design, this suggests that it is more likely an intentional decision on the part of the participant to maintain the same estimate.

We did not give additional assistance to participants who chose to realign their estimate on one trial to their estimate on their last. Because participants are given the additional task of precisely realigning their estimate on one trial with their last estimate, this introduces a source of error. Participants who intend to maintain the same estimate may miss their last estimate by some margin. This will cause deviations in response from trial to trial even while the participant's true estimate remains the same. Thus, participants who are asked to respond every trial may change their estimates of the Bernoulli parameter by small amounts more often than do participants given a default response. These additional adjustments however would be due to added noise around response-maintenance rather than reflecting true new updating events.

New updating events ought to make the most recent sample more likely. Unlike noise, updates to one's estimate are presumably going to be in a direction consistent with the most recent sample rather than inconsistent with it. If participants given no default response genuinely update more often, they would make more sample-consistent adjustments to their recent estimates than do participants given a default response. In contrast, if it is

noise added to participant responses, any new adjustments would be just as likely to be sample-inconsistent as sample-consistent.

Thus, while the participants who are forced to respond anew every trial may make more adjustments, whether these adjustments reflect true intended updating events or noise can be determined by whether those adjustments tend to be sample-consistent. If participants do make more adjustments which are in-line with recent events, this will indicate that an effort threshold has been removed and participants are genuinely updating more often. If participants make more adjustments but those adjustments are just as likely to be sample-inconsistent as they are to be sample-consistent, this would suggest that they are intending to maintain their response. If response maintenance is intended but deviations occur, this would still support response-maintenance as a natural component of mental model updating. Indeed, if step-hold patterns persist in human estimates when no default response is provided, then this supports the conclusion that a thresholded updating process is not due to effort but is more intrinsic to human probability estimation methods.

Chapter 2

Methods

2.1 Demographics

Participants were recruited from the University of Waterloo student body and offered research participation credit as compensation for their participation. All participants gave informed consent and the study was cleared by the Universitys Office of Research Ethics (ORE 42844).

The study was conducted online. The experiment used a University of Waterloo server. It was coded in PHP and Javascript and used elements from the JsPsych library [25]. A total of 229 participants enrolled; 56 participants left before experiment completion. 173 (111 female, 55 Male, 6 no response, and 1 non-binary) completed the task. Only data from participants who completed the task was used for analyses. Mean age was 20.42 (SD = 2.2), with a range of to 17 to 33.

2.2 Preliminaries and Practice Trials.

Two independent data collections were run. Both had the same procedures with some minor variations noted below. Participants were first informed that the task was expected to take one hour, and this was followed by an online informed consent procedure. Next participants responded to some demographic questions. A set of practice trials began the experimental procedure. After completing the practice trials and the full experimental task, participants were asked about their impressions of the task and debriefed.

2.3 Experimental Procedures

The task was the same for both data cohorts, but for the second collection we shortened the instructions and included a progress-bar. We also gave feedback on the practice trials in the second cohort; if participants performed too poorly on the first five trials, they were asked to start again. Since the experimental protocol for the task itself did not change, for the following analyses, the separate samples are collapsed and both cohorts are used.

2.4 Task Details

The trial display (see Figure 2.1) consisted of a box in the middle of the screen which contained 25 dots. Before a participant had made any response, the colour of the dots was greyed out. Below the box, a slider recorded the participant's estimate of red to blue dots. After the first click the portion of the slider-bar which represented the proportion of red dots became red, and the opposite portion of the slider became blue. The proportion of the dots in the urn also responded to slider movements to show a matched proportion of red and blue.

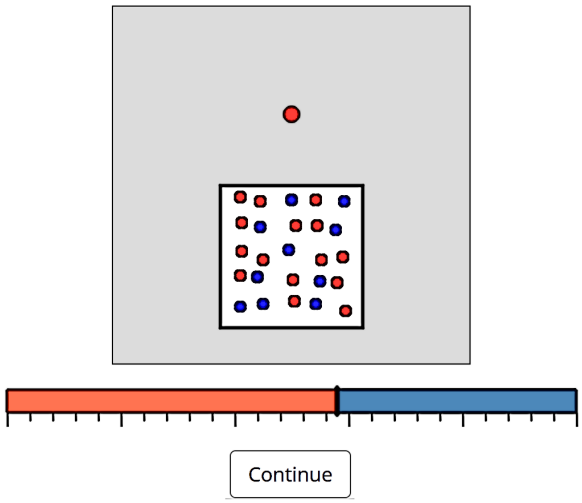

Figure 2.1: Experimental display
On every trial a red or blue dot floated out of a box. Participants were asked to adjust the slider to represent their estimate of the proportion of red to blue dots in the box.

2.4.1 Practice Trials

Participants were given 30 practice trials to practice placing their mark on the slider. For practice trials, a grey line was drawn on the slider and participants were asked to click on the line. The position of this target in the practice trials was randomly chosen. The precision of participants' clicks was used as a measure of their tolerance for slight errors in accuracy for the subsequent task.

2.4.2 Experiment Trials and Conditions

There were 999 trials, split up into 3 blocks of 333 trials. After 333 trials, a screen would appear informing the participant that they were on a break.

On each trial participants witnessed a coloured dot float out of the urn. The slider below the urn represented their estimate of the proportion of colored dots in the urn for that trial. Changes in estimates required adjusting the slider position.

For those trials where adjustments were made, adjustments were classified by their direction and magnitude. Adjustment direction could be either sample-consistent or sample-inconsistent. For a sample-consistent adjustment, a participant changed their Bernoulli estimate to make the most recently observed dot colour more likely. For a sample-inconsistent adjustment, a participant moved their slider in the opposite direction of the dot colour. For instance, if the most recent dot had been red and the participant *decreased* their estimate of red this would be a "sample-inconsistent" adjustment. Adjustments of zero size correspond to response maintenance and are not considered 'adjustments' for our purposes. As well as a direction, any adjustment was also associated with a magnitude, referring to the number of slider positions over which the participant moved their estimate.

The slider had a total of 101 possible settings representing the range of proportions of coloured dots from 0 to 100 percent. Once a participant was satisfied with their estimate, they would click the 'Continue' button, below the slider, to move on to the next trial.

The colour of the dot that floated out from the urn could be red or blue. The probability of either color was randomly determined by selecting uniformly from 0 to 1. This probability changed in an unannounced fashion with the length of the run before the next change point chosen uniformly from 1 to 100.

Participants were in one of two conditions: 'Automatic' or 'Manual'. As in previous research using this task (e.g.[4, 24, 6, 5]) participants in the Automatic condition were given their last response as their default response on each trial. For these participants, at

the onset of a new trial, the position of the slider was simply maintained from the previous trial. Participants could simply click the 'Continue' button to move onto the next trial. In contrast, participants in the Manual group were given no default response at all and were asked to respond anew on every trial. They were presented a grey slider with a mark at the location of their prior estimate. They were explicitly informed that to maintain their most recent estimate, they needed to click on this line. If participants attempted to click the 'Continue' button without adjusting the slider, they were given an error message in red text which read 'Please indicate your response on the slider'.

2.4.3 Exclusion

Only those participants who performed the task correctly and well, following the given instructions, were to be included in the analysis. Previous research has already established that participants are overwhelmingly capable of performing very well on this task (see [4]). The present study was concerned with the way these accurate estimates are generated, specifically, how quickly participants incorporate new data to update their estimates of the probability. Thus, exclusion criteria were directed towards removing participants who misunderstood the task, were distracted during the task, who did not show sufficient vigilance to perform well on the task, or those who performed the task poorly for other reasons. Whether participants correctly followed instructions was measured by their accuracy on the task and the length of time it took them to complete the task.

For accuracy, the unstandardized beta coefficient was used, predicting the true Bernoulli parameter from the participants estimates of the Bernoulli parameter. This gave a measure of how close a participant's estimates of the Bernoulli parameter were to the true Bernoulli parameter.

Accuracy was segmented into performance per block and exclusion criteria concentrated on these segments. The segmentation was due to a bifurcation of participants in later blocks into good and poor performers. As can be seen in Figure 2.2, a cluster of participants who do not appear to attempt to do the task emerges by the third block. Pictured is an apparently bimodal distribution, with one cluster of participants with near zero correlation with the true Bernoulli parameter and another cluster of participants with higher sample-consistent correlations.

Such emerging poor performance was attributed to a fatigue effect, where some participants over the course of the task became less engaged with the assigned task. In order for a participant to be included in the analysis, their estimates of the Bernoulli parameter would need to significantly predict the true Bernoulli parameter across all three blocks.

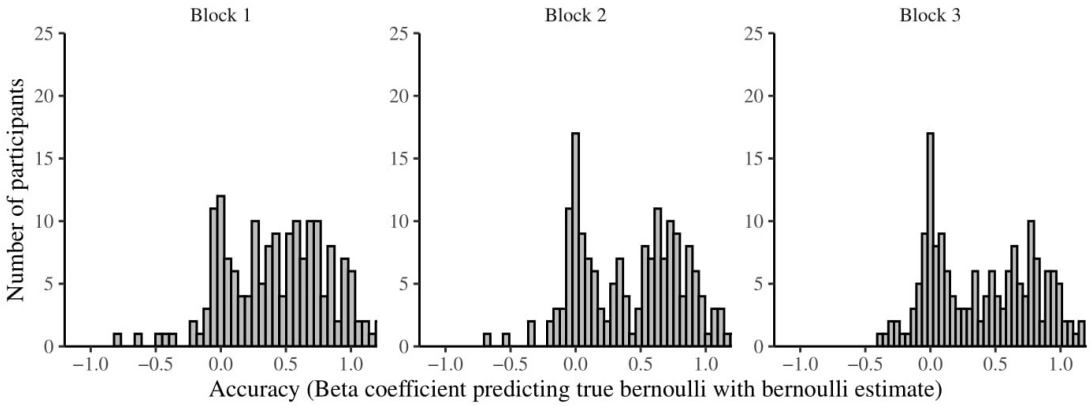

Figure 2.2: Histograms of participant accuracy separated by block.

The critical accuracy beta coefficient, .108, corresponded to a significance test with alpha value of 0.05. Participants whose accuracy fell below this critical value on any of the three blocks were excluded. Of the 84 participants in the Manual group, 44 were excluded on these criteria. Of the 89 Automatic condition participants, 32 were excluded. The total number participants removed for failing to meet the accuracy criteria was thus 76.

The length of time it took participants to complete the task ranged from 3.25 minutes to 22.57 hours. Outliers were not identified using the standard deviation of task duration because this statistic would be disproportionately pulled upwards by extremely long task times. Instead, quartiles, which are much less sensitive to outliers, were used. Participants who took longer than 1.5 times the inter-quartile range (IQR) over the 3rd quartile were excluded. A total 10 participants took more than 91.71 minutes to complete the task. Of these participants, 5 had already been excluded for failing to meet accuracy criteria, so only an additional 5 participants were removed on the account of taking too long to complete the task. The number of participants in the final sample was thus 92. Task duration ranged from 5.17 minutes to 85.41 minutes.

Chapter 3

Results

3.1 Survey Responses

Participants completed the task from home unsupervised, and thus it is possible that they were discretely engaged in other activities while they were doing the task. At task completion, participants were asked to type in a box any secondary activities they had been engaging with during their participation. Of the 173 participants who completed the experiment, 83 participants reported engaging with a secondary activity during the task. Of these 83 participants, 41 reported listening to music, 14 reported engaging with a second media (such as listening to a podcast or watching a television show), 4 reported engaging in a social activity (such as talking to a friend), 7 reported attending to phone notifications, and 7 reported engaging in a miscellaneous activity (such as eating dinner or petting a cat). Whether a participant was engaged in another activity did not predict accuracy $F(1,171) = 0.112$, p = 0.739, task duration $F(1,171) = 0.91$ 0.341, nor number of times the task window lost focus $F(1,171) = 0.459$, p = 0.499. In the final sample, within the 92 included participants, 45 reported engaging with another activity during the task. Of these participants, 18 reported listening to music, 7 reported engaging with a second media, 2 reported engaging in a social activity (such as talking to a friend), 4 reported attending to phone notifications, and 5 reported engaging in a miscellaneous activity

3.2 Practice Trials

Of the 173 participants who finished the task, 44 perfectly aligned their response with the provided practice line and thus made no errors on the practice trials. Within the subgroup of the 92 participants who met inclusion criteria, 32 made no errors on the practice trials. Within this group of included participants, the median number of errors was 2 out of the total 30 practice trials; the range of number of errors was from 0 to all 30 of the trials. Participants can easily maintain their prior response when they choose to.

Only the Manual group was asked to re-click the slider on every trial, and thus their task was more comparable to practice trials. The Manual group took significantly longer on practice trials ($t(30)$=8.05, p <0.001) and corrected their adjustments significantly more often(26%) than in the full task (M=9 percent of trials; t(30)=3.56, p<0.001).

3.3 Effects of Condition on Task Performance

Of primary concern was the question of whether providing participants a default response introduced an effort threshold preventing frequent updating. This was first tested by comparing the number of adjustments made by the Automatic group to the number of total adjustments made by the Manual group. If the default response introduced an effort threshold to the Automatic group, we would expect to see the Manual group make more adjustments than the Automatic group. Consistent with such an effect, a t-test established that condition significantly changed frequency of response maintenance: Manual participants adjusted their responses away from the previous trial significantly more often than did Automatic participants ($t(90) = 7.09$, $p < 0.001$; Figure 3.1). Average number of adjustments for the Manual group was 705.03 trials out of the full 999 trials. For the Automatic group it was 320.96.

Given that Manual participants make more adjustments, it was next investigated whether these additional adjustments were in-line with recent evidence.

For each adjustment size of the report slider we computed the average number of adjustments made by the Automatic group. This created a vector representing the Automatic group's average number of adjustments per each adjustment size. To isolate the differences between groups in their distribution of adjustments, for each participant in the Manual group for each adjustment size, we subtracted the Automatic group average number of adjustments. These residual vectors for each Manual participant represent adjustments

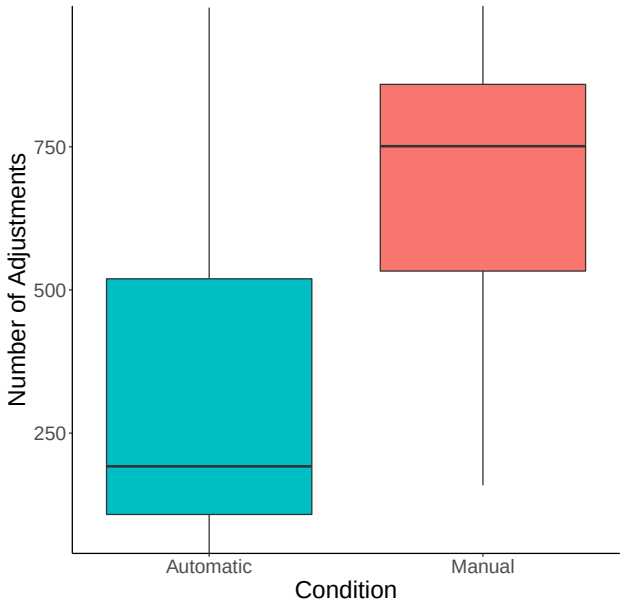

Figure 3.1: Boxplot showing group differences in total number of adjustments

Manual participants made of more than the average Automatic participant and are referred to as 'added adjustments'.

To first determine whether these added adjustments were in-line with recent events, the number of added adjustments was collapsed over adjustment size, preserving only adjustment direction. A simple paired t-test determined that, when Manual participants make more adjustments, those adjustments were significantly more likely to be sample-consistent than sample-inconsistent $t(36) = 2.87$, $p = 0.007$.

A Chi square test determined that the proportion of adjustments being sample-inconsistent was unusually high for added adjustments compared to regular adjustments $X(1) = 58.72$, $p < 0.001$. For Automatic participants, the percent of adjustments made over the course of the task which were sample-inconsistent was 10%. For Manual participant added adjustments this was 35%.

Next, it was investigated how much of new Manual participant responses could be attributed to either the noise or true new updating hypotheses.

We fit a mixture model to the number of Manual participant added adjustments and compared this mixture model to a model with only a single Gaussian random variable. Non-linear model fitting was done using R's nls function [26]. The single Gaussian model was fit

using a mean fixed at zero. The mixture model was composed of two Gaussian distributions. For one Gaussian, the mean was fixed at zero, for the other Gaussian the mean was allowed to vary and entered as a free parameter. Both fit models are shown in Figure 3.2.The mixture model provided a substantially better fit to the Manual group's residual added adjustments than did the single Gaussian model. The AIC decrease was 200.58. This evidence supports the hypothesis that the adjustments which Manual participants made more of consisted of two distributions: one of noise around intended response maintenance, and one of increased disposition to make sample-consistent adjustments in-line with recent evidence.

Given the best-fit mixture model, the area under the distribution represents the total number more adjustments the Manual group made over the Automatic group. The area considered excludes adjustments of zero size, because these would not be considered adjustments. Of the total area of 321.26 added adjustments, 38 % (124.05) was contributed by the Gaussian distribution representing noise, and 61.40 % (197.31) was contributed by the Gaussian distribution representing true intended updating events. The mean of the Gaussian distribution representing true updating events was 1.2. This represents that, those updating events which the Manual group made more of, on average, were 1.2 slider positions over from their previous response.

3.4 Individual Variability within Conditions

While these marginal effects of group on number of adjustments are significant, there were also large within-group differences. Particularly tellingly, some Manual participants made fewer adjustments than did the average Automatic condition participant. The result of this is that their number of 'added' adjustments was negative. This reflects them having actually made fewer adjustments. Individual total number of added adjustments in the Manual group ranged from -170.18 to 666.82. Figure 3.3 shows four of such Manual participants, each maintaining their previous estimate more than 73.08% of the time. Two Manual participants who made the fewest number of adjustments per their group are not shown. These participants each made zero and two total adjustments over the full task.

Figure 3.4 shows analogously meticulous Automatic participants, who made more adjustments than the average Manual participant.

The range of the Automatic total number of adjustments was 41 to 993. For the Manual group the range was 159 to 996.

Figure 3.5 displays, by group, the range of number of adjustments over the task and the

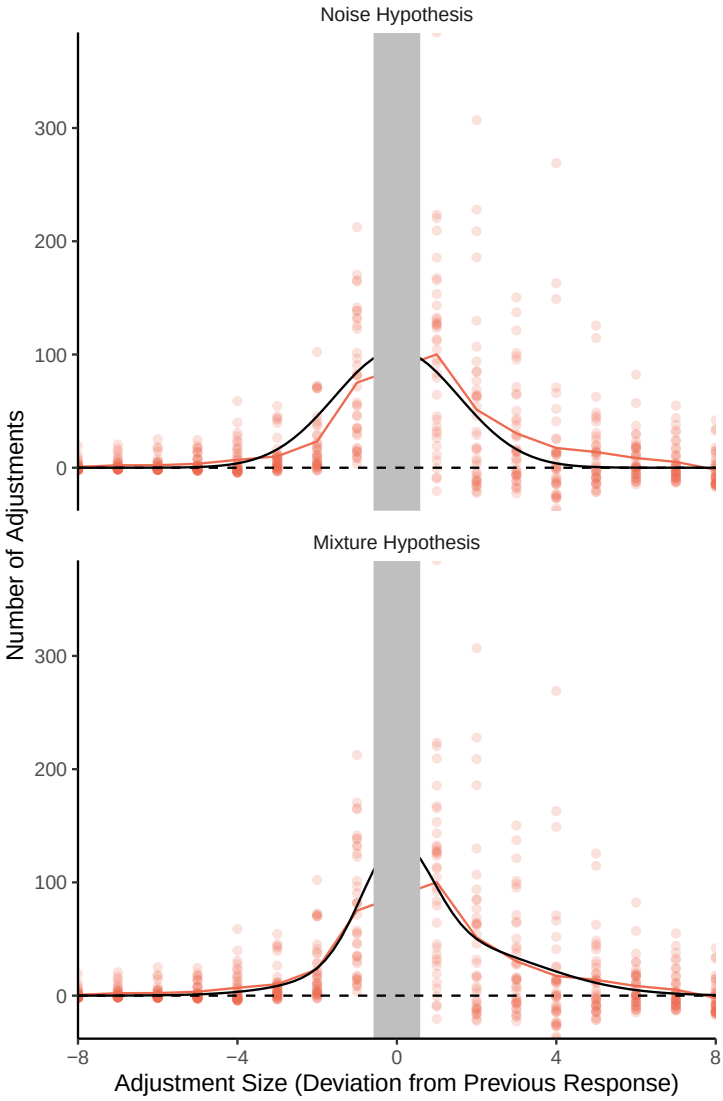

Figure 3.2: Compared model fits for Manual participant added adjustment distributions Manual participant added adjustments over the course of the task. Adjustment size is delineated in slider units, where an adjustment size of one corresponds to a shift in one's Bernoulli parameter estimate by one of the 100 slider positions. Each red dot represents the added adjustments of one Manual participant at a particular adjustment size. The red line represents the average for the entire Manual group. The black line represents the best-fit curve for each of the two models that were fit to participant data: the just noise model and the mixture model. The vertical grey line blocks the section where no data was available for model fitting. Adjustments of zero size do not qualify as added adjustments and would not be valid to compute.

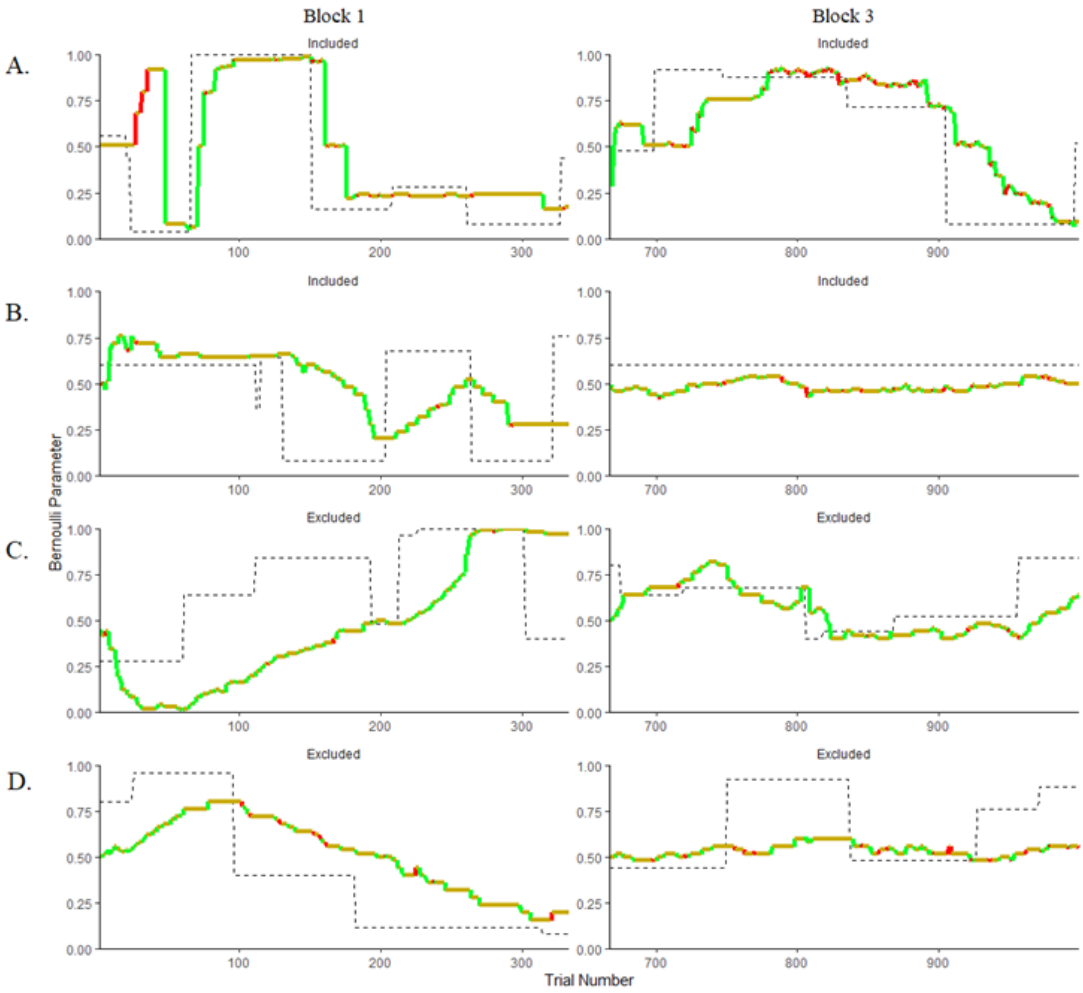

Figure 3.3: Performance over task of Manual participants who made the fewest adjustments Four Manual participants with the highest frequencies of response maintenance. Each row corresponds to a participant, and each column corresponds to a block (left is Block 1, right is Block 3). On the y-axis is the Bernoulli parameter, and on the x-axis is trial number. The dashed line corresponds to the true Bernoulli parameter on these trials. The coloured line corresponds to the participant's estimates on these trials. Sample-consistent adjustments (in-line with recent evidence) are green; sample-inconsistent adjustments are in red. Response maintenance (no adjustment) is in yellow.

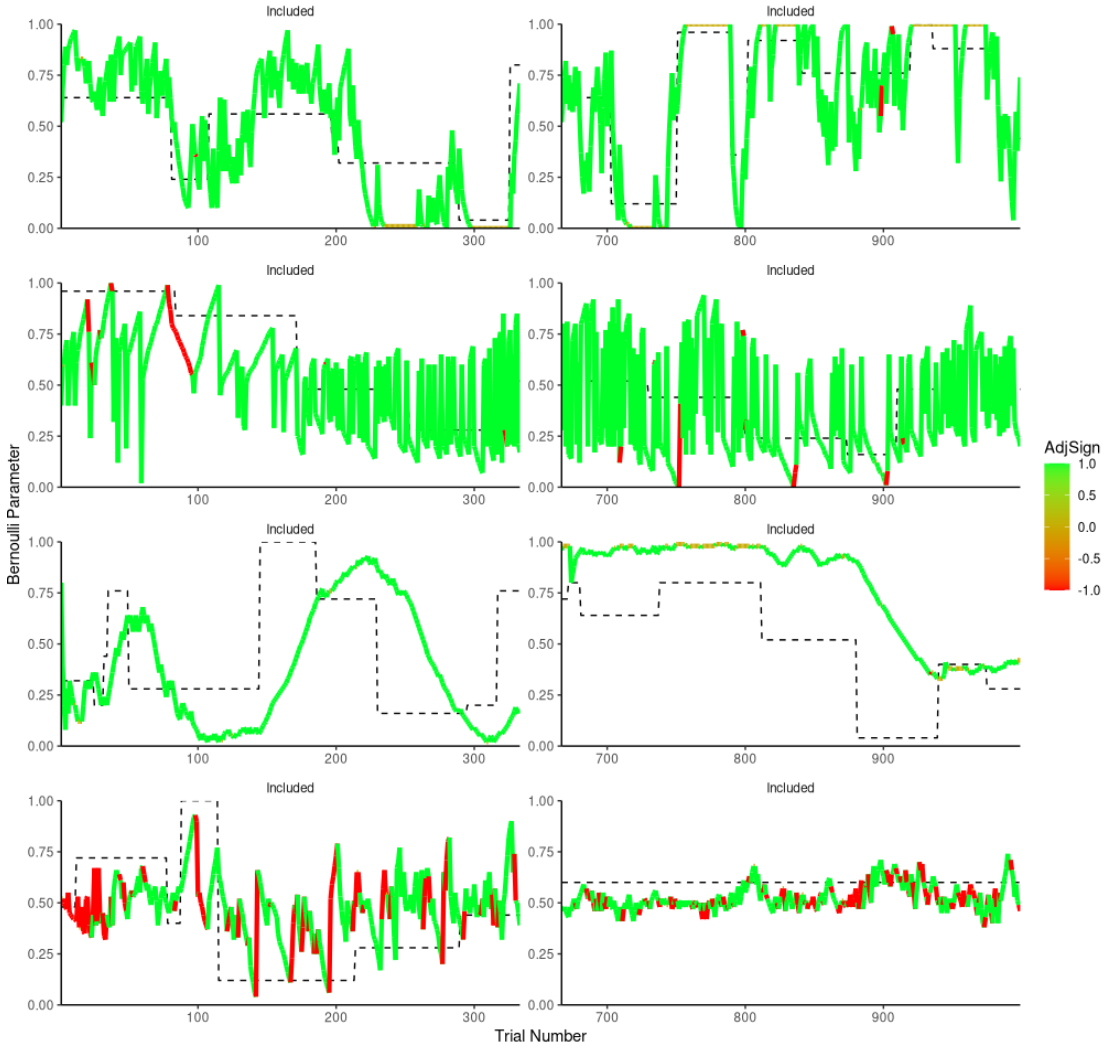

Figure 3.4: Performance over task of Automatic participants who made the most adjustments

Four Automatic participants, meeting inclusion criteria, with the highest frequencies of adjustments. Each row corresponds to a participant, and each column corresponds to a block (left is Block 1, right is Block 3). On the y-axis is the Bernoulli parameter, and on the x-axis is trial number. The dashed line corresponds to the true Bernoulli parameter on these trials. The coloured line corresponds to the participant's estimates on these trials. Sample-consistent adjustments (in-line with recent evidence) are green; sample-inconsistent adjustments are in red. Response maintenance (no adjustment) is in yellow.

extent to which these adjustments were sample-consistent. In the Automatic group, there is a clear strict correspondence between the adjustment rates. When Automatic participants make more adjustments, this corresponds to them making more sample-consistent adjustments. This is represented by how tightly the Automatic group points align with the diagonal line across the plot. This relationship contrasts the pattern seen in the Manual participants. Consistent with the noise added to their responses, Manual participants who made more adjustments often correspondingly also made substantially more sample-inconsistent adjustments.

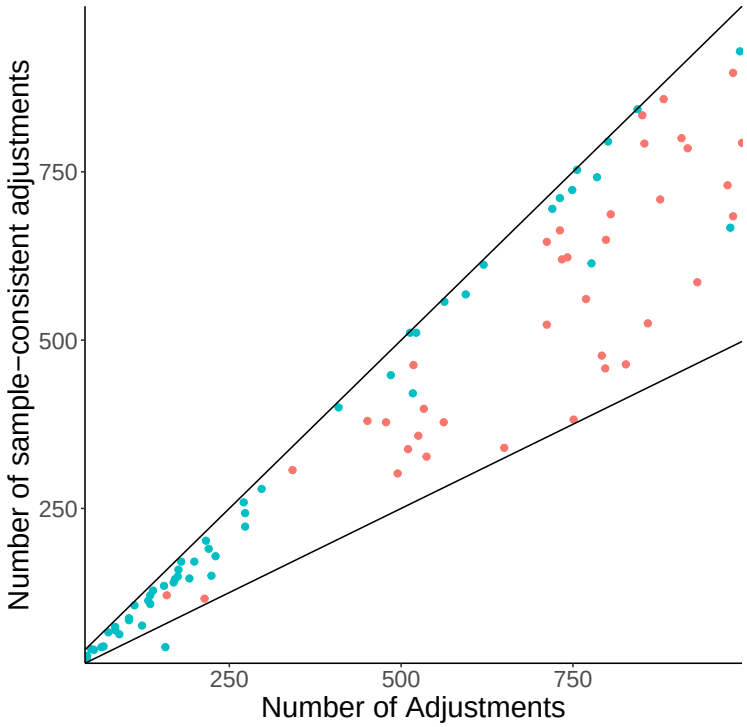

Figure 3.5: Number of sample-consistent adjustments by number of total adjustments
Each participant's number of sample-consistent adjustments plotted against their total number of adjustments. Red dots represent Manual participants; blue dots represent Automatic participants.

To perform statistical analyses, number of sample-consistent adjustments was not used due to its dependence on total number of adjustments. Instead, proportion of sample-consistent adjustments was used. Figure 3.6 displays the unexpected non-linear relationship visible in the Automatic participant data. For only the Automatic group, there is a

significant correlation between number of adjustments and proportion of those adjustments being sample-consistent, $F(1, 48) = 20.25$, $p < 0.001$, $R^2 = 0.296$. This correlation is not significant for the Manual group, $F(1, 30) = 1.23$, $p = 0.276$. To test the non-linearity of the relationship, adding the log transform of the number of adjustments to the regression equation significantly improved the fit of the model, $F(1,47) = 16.66$, $p < 0.001$, $R^2 = 0.476$. When Automatic participants make few adjustments, those adjustments may be just as likely to be sample-consistent as sample-inconsistent. However, as Automatic participants make more adjustments, it rapidly becomes the case that those adjustments are almost always sample-consistent.

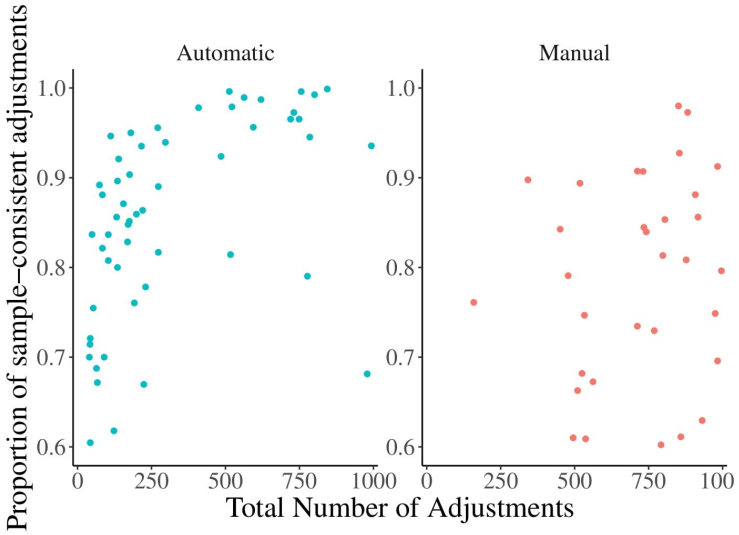

Figure 3.6: Proportion of sample-consistent adjustments by number of total adjustments Each participant's number of total adjustments over the task plotted against the percentage of these responses being sample-consistent. Red dots represent Manual participants; blue dots represent Automatic participants.

Chapter 4

Discussion

4.1 Study Overview

We were interested in whether response maintenance patterns in previous research could be explained by the extra effort it takes to update one's response. We compared two conditions: one where participants were given their last response as default and another where they were given no default response at all. If it were the case that providing default responses introduces a threshold to response updating, then participants in the Manual group should update their responses more often than the Automatic group. Consistent with this effect, Manual group participants on average made more adjustments to their previous estimate than did participants in the Automatic group.

However, these added adjustments are not a pure measure of increased updating events. Our experimental design introduced an additional task for the Manual group. Only Manual participants had to precisely re-align their slider with a grey bar in order to maintain an exact estimate from trial to trial. Performance in the practice trials indicated that it was certainly possible for Manual participants to maintain a response if they so intended. However, it is also possible that this additional task added noisy deviations around intended response maintenance.

Two hypotheses were distinguished to explain added adjustments in the Manual group. According to the noise hypothesis, when Manual participants intended to maintain the same estimate, sometimes they tolerated some deviation from one trial's estimates to the next. This produced differences in estimates that did not mark true intention to update a response. In contrast, according to the true new updates hypothesis, an effort threshold

prevented Automatic participants from making as many adjustments to their estimate as often as their mental models changed. When this threshold was removed, response maintenance became no easier than response updating for Manual participants. Manual participants made new adjustments which corresponded to true intended updating events.

The two hypotheses were distinguished by the proportion of adjustments which were sample-consistent or sample-inconsistent. If added adjustments were just noise, they should be equally likely to be in-line with recent evidence as contrary to it. If added adjustments marked true new updating events, then they should be sample-consistent. Our data support both hypotheses partly explaining Manual participant added adjustments. There was both added noise and new updating events.

Consistent with the true new updating events hypothesis, added adjustments in the Manual group were more likely to be sample-consistent than sample-inconsistent. Consistent with the added noise hypothesis, sample-inconsistent events were more common in Manual participant added adjustments than in Automatic regular adjustments.

To estimate the proportion of adjustments contributed by either hypothesis, a mixture model was fit to Manual participant added adjustments. To represent the hypothesis that Manual participant added adjustments were only noise and not true updating events, a Gaussian distribution with a mean fixed at zero was fit to the Manual participant's added adjustments. The mean fixed at zero forced the best-fit Gaussian to estimate that adjustments were just as likely to be in-line with or contrary to recent events. In contrast to the added noise hypothesis, the alternative hypothesis was that the Manual group added adjustments represented more frequent updating. This hypothesis was also represented as a Gaussian distribution, but with the mean entered as a free-parameter and allowed to vary away from zero and thus be sample-consistent. A mixture model composed of both the noise and the updating hypothesis represented the hypothesis that added adjustments reflect both noise and true new updates.

Model fit was substantially improved between the just-noise model and the mixture model. This suggests that Manual participants both had noise added to their intended response maintenance and that they genuinely intended to update their responses more often. Interpreting the statistics extracted from this mixture model can be meaningful, but caution should be exercised when it comes to accepting the particular numeric quantities. According to the best fit mixture model, 40% of new updates were noise and 60% of new updates were true new updating events. The average adjustment size of these new updates was estimated to be 1.2 slider-units, of the total available 100 slider positions. The small size of these adjustments is consistent with the removed effort threshold hypothesis. Adjustments at this scale would be negligible if a participant hurried through the task.

The previous discussion covers average group differences. However, the effect of the condition manipulation was certainly not uniform across participants. Within groups there was a tremendous amount of individual variation in task strategy. Both the Automatic and Manual group number of total adjustments spanned almost the full range of possible number of trials. There are particularly striking cases of particularly meticulous participants. Before exclusions, six Manual participants updated their estimates less often than did the average Automatic participant. Two of these participants did not update their response at all and are not considered to have shown meticulous response maintenance. Because these two participants maintained the estimate at 50% red 50% blue throughout the entire task, this behavior is taken as indication that these participants had a firm belief in their complete and unshakable uncertainty.

Four remaining Manual participants updated less often than Automatic participants. This Manual participant meticulous response maintenance is taken as evidence that, despite it being more effortful, these participants truly intended to maintain the same estimate of the probability over series of trials. This is despite possible perturbations to the estimate from unpredictable events driven by the stochasticity of the process. This suggests that thresholds introduced by previous task design does not fully explain why participants demonstrate step-hold patterns in their probability estimates. Response maintenance can appear spontaneously and can be apparently quite intentional and effortful.

Automatic participant responses were found to be more clean than Manual participant responses. Plotting Automatic participant adjustments against sample-consistent adjustments more cleanly shows the full continuous range of the possible updating frequencies from trial-by-trial to threshold updating.

An unexpected non-linear effect was found only in the Automatic group regarding the proportion of their adjustments being sample-consistent. When Automatic participants made few adjustments, those adjustments may be just as likely to be sample-consistent as sample-inconsistent. When adjustments are rare, it is more sporadic whether those adjustments correspond to incorporating the most recent event. However, when Automatic participants made any more than very few adjustments, those adjustments tended to be over 90% sample-consistent.

Sample-inconsistent adjustments may be surprising in the Automatic group. The Automatic group is never forced to respond, and so we would expect that Automatic participant adjustments mark true intent to update their estimates of the probability. If recent evidence is being used to update probability estimates, the most recent sample should be what determines the direction of the update. However, sample-inconsistent adjustments are a common finding in previous versions of this task. Forsgren et al. estimate that the

proportion of sample-inconsistent updating is around 25%.

In our study, we found that sample-inconsistent adjustment frequency was related to number of total adjustments made by Automatic participants. We explain the relationship as there being a tighter temporal correspondence between mental model updating and slider response updating in Automatic participants who update often. When adjustments are infrequent in the Automatic group, adjustments may be delayed. Prompter response time would be expected for participants whose estimates are more swayed by recent events. This explanation is consistent with a hypothesis offered by Forsgren et al [24]. By their account, sample-inconsistent responses are due to a variable threshold preventing prompt updating. On one trial where the threshold is high, evidence may push the estimate to change by a substantial amount, but no adjustment is made. If, by chance, the threshold on the next trial is lower, even while the evidence pushes the estimate in another direction, the substantial shift in estimate from the previous trial still determines the adjustment direction. Thus, adjustments inconsistent with the most recent evidence are due to a time-delay in updating caused by a variable threshold.

The large amount of individual variation within groups speaks to a continuity of updating strategies which ranges from immediate updating following recent evidence to only intermittent stepped updating following accumulated evidence. Thus, regardless of specific learning algorithm, response maintenance seems to reflect true mental model updating, rather than is it exclusively explained by there being a motor cost to response updating. The presence of participants who update nearly every trial suggests that the mental model updating mechanism at least allows for immediate updating, and so it cannot be the case that the updating mechanism precludes immediate incorporation of evidence to mental models.

4.2 Explaining Spontaneous Response Maintenance

Some participants in the Manual condition showed spontaneous response maintenance, despite it being the more effortful strategy. This step-hold pattern cannot be explained by increased effort to update a response. Rather, there may be other reasons that updating is infrequent. One possibility was proposed by both Gallistel. et al [4] and Khaw et al. [6]. There may be a cognitive cost to updating. It may be computationally expensive to compute a new estimate of the probability. Adding a threshold procedure turns probability estimation to a two-step process; the updating subroutine is only engaged when a participant makes the binary decision to update on that trial. Thus, the cost of updating is made exclusive to only those trials an update occurs.

This explanation only holds if mental model updating can be delayed and does not occur with each sample. This rules it out as an explanation for step-hold updating if the learning algorithm used to generate estimates is a trial-by-trial model. However, Forsgren et al. [24] found that the best fit to their participant estimates was a indeed two-step trial-by-trial updating model. An explanation for intermittent updating is still needed if the updating subroutine is not computationally expensive and occurs with each sample anyway.

One possibility is that a fixed estimate produces a more stable reference point to compare incoming statistics of new samples to. As an estimate of a probability stabilizes over accumulating evidence, fixing an estimate over several trials may prevent the estimate from drifting towards a new mean if the stochastic process suddenly changes. A fixed reference point for hypothesis comparison may make differences between first-order stochastic processes more discriminable.

Another potential benefit to response maintenance was proposed by Gallistel et al. [4] and was reviewed earlier. Estimating runs over which a stochastic process remained fixed allows for more efficient storage of information that is not tied to every single data point. This strategy is good for retrospective inference. Taking the mean of a set of samples is less noisy than the samples themselves. Remembering a set of runs with estimated stochastic processes will allow a person to make inferences about the general pattern observed as it unfolded over time.

There are thus benefits to adding a threshold procedure to prevent updating. This is whether or not the generating of new estimates is a subroutine only engaged when the threshold is met.

4.3 Relating Response Maintenance to Hypothesis Testing

Suspended updating by the above account is a consequence of hypothesis testing. In suspended updating, different explanatory hypotheses are discretely switched between resulting in runs of a maintained estimate. This response maintenance is not, to be clear, an inevitable consequence of hypothesis testing.

Thus far, trial-by-trial updating and hypothesis testing have been presented as if they are mutually contradictory. After all, if every sample is used to update a response, then samples are not being accumulated. Much like trial-by-trial updating can be made into

change-point detection given a threshold procedure, change-point detection more closely resembles trial-by-trial updating if the threshold procedure is removed. Adjustments can occur every trial. This is because new samples still contain information about the hidden stochastic process, even if it that process is assumed to be fixed. Samples can be used to amend hypotheses by minor amounts even while the model estimates that the underlying probability has not changed. Volatility in estimates, given a fixed change-point location is especially true in the early stages where a new change-point has just been inserted.

This becomes apparent when the threshold procedure is removed from the IIAB model and adjustments to probabilities begin to occur at every trial. This is because the updating subroutine uses optimal estimates of the probability, regardless of whether a change-point was added, removed, shifted, or left alone. Optimal estimates of the probability are going to incorporate the most recent evidence and thus be partially altered by it.

Hypothesis testing is thus separable from whether updating occurs with each sample or not. As identified by Forsgren et al [24], the main difference between the delta-rule model and the IIAB model is the weight assigned evidence. In the delta-rule model, more weight is assigned to recent evidence, whereas in the IIAB model, all data since the last change-point is given equal weight.

The increased optimality of the IIAB model probability estimates requires extensive memory. In the task Gallistel et al. [4] assigned to participants, the average run-length was 200 trials. The required average memory store for the IIAB model updating subroutine was thus 400 trials. This memory requirement is sensitive to the precise location of each dot in the time series. The advantage of this increased optimality is questionable; delta-rule model estimates are so near optimal they are hardly discriminable from optimal Bayesian change-point detection estimates [27].

This kind of trial-by-trial variation in change-point detection model estimates is distinct from true trial-by-trial updating in one particularly distinctive way. Models which use change-points to estimate probability will probably weigh incoming samples less and less the further a sample is from the estimated change-point location. An optimal change-point detection algorithm will want to use as much gathered evidence as possible to make estimates of current probabilities and will not be pulled around by recent events if it estimates that the generative process has not changed. It would thus be expected that over long runs, adjustment size will decrease as estimates converge. This is unlike a delta-rule model with a fixed learning rate, where adjustments would depend only on error. Empirical data support the change-point detection method here. As previously discussed, variable adjustment strength, such as implied by change-point detection algorithms is consistent with the finding that participant learning rates decrease as the most likely run-length

increases [8, 27].

This does not however indicate that change-point detection algorithms best explain participant estimates. A delta-model algorithm with a variable learning rate, for instance, would also explain this decrease in adjustment size [8]. Furthermore, an empirical finding which draws into question the use of optimal change-point detection algorithms is learning rates are certainly not optimal [7]. In particular, once the optimal Bayesian observer estimates stabilize via converging to the true probability over a run, human estimates remain more volatile citecarrabin2021. Human estimates are more subject to change given recent evidence than is optimal given the accumulated evidence thus far. This may be explained by a residual epistemic uncertainty in human estimates. If the inference process is imperfect and known to produce unreliable estimates, then there will always be an advantage to relying more heavily on recent events. This exaggerated uncertainty given the weight of evidence is consistent with previous research finding that, while human estimates are near accurate, the confidence that people claim confidence in those estimates is severely underweighted [28]. Weighting events by their recency will ensure that a model is kept relevant. This is a natural advantage offered by the delta-rule model. It is also possible to adjust change-point detection algorithms to weigh recent events more heavily if estimates of epistemic uncertainty are part of the equation.

Overall, there being response maintenance in participant responses is not sufficient to determine whether change-point detection and hypothesis testing occurs.

4.4 Expanding the Notion of Hypothesis Testing

Hypothesis testing as a probability estimation tool has been presented quite simply and has not been fully explored. It has a broader role to play in mental model updating which requires elucidation. Thus far, hypothesis testing has been related to two-step processes for mental model updating. In two-step mental model updating, hypothesis testing occurs at the first stage where it is detected whether a change in probability is likely to have occurred. This simple use-case severely limits the potential offered exclusively by incorporating hypothesis testing into mental model updating. This potential is only touched upon by Gallistel et al.'s treatment of the method.

A simple point about accumulating information for hypothesis testing should first be made. In the IIAB model, using the KL divergence, the estimated probability distribution is compared to the observed probability distribution. This observed probability distribution is the data accumulated since the last purported change-point. However, this means that as

long as no new change-point is inserted, the influence of recent samples on the estimated probability distribution decreases with distance from the change-point. Therefore, the effect of samples declines rapidly over time, making the estimated distribution increasingly slow to pick up potential sudden changes in probability. This is an issue if the model is meant to quickly and optimally detect changes in probability.

In the IIAB model, the decreased weight of new evidence on the observed distribution is counterweighted by a second measure. E is calculated by taking the product of the KL divergence and the number of observations made since the most recent estimated change-point. This means that the larger the sample size, the more likely the model is to detect a change. However, while the change-point threshold may then inevitably be reached, this may have less to do with recent data, and more to do with the number of samples that have passed. This is does not then represent a pure accumulation of model error. For the model to be sensitive to sudden changes, a leaky integration model, such as is offered by the delta-updating rule may fare better. This will discount information stretching back in time before the probability changed. Another alternative, offered by Forsgren et al. [24], is using drift diffusion as the threshold procedure.

4.4.1 Incremental Single Model Adjustment versus Model Comparison

There is a hidden premise in hypothesis testing which requires uncovering. Hypothesis testing using a threshold procedure is not as simple as determining whether data are sufficiently unlikely given a working model. This is the point delicately made by Griffiths and Tenenbaum [29]. Hypothesis testing requires there being an alternative model which better explains the data than the current model. This requires at least simultaneous entertainment of two hypotheses at any one time. Comparing the relative plausibility of more than one model resembles model comparison in a posterior distribution. That is, even with trial-by-trial updating, a single estimate of the probability does not store all the information relevant to the stochastic process. More than one explanation must be held simultaneously in mind in order to detect that there is a better model to be moved to.

This implicated necessary model comparison is quite profound. Its necessity has been obscured due to the simplicity of the doubly stochastic Bernoulli process. In the Bernoulli estimation task, updating estimates is very straightforward and follows a single trajectory. A better estimate which weighs the most recent event more strongly is derived simply by moving the current estimate of the probability more towards one or the other extreme. The success of the delta-model in explaining human estimates of the probability may have more

to do with the simplicity of the assigned task than its generalizable applicability. Heilbron and Meyneil [30] argue that, in these tasks, flat delta-model estimates which conveniently forget the past and rely on recent data are too close to optimal Bayesian learning to truly discriminate between the two strategies. The delta-model is able to generate estimates from its previous estimates so effectively because, unless the current estimate is at the extreme, an alternative better fitting model is always implied by each individual sample.

But deriving that there exists a better fitting model and thus the current model no longer applies is not typically so straightforward. When the model is more complex, a single sample, for example, cannot be assumed to imply the new better-fitting mental model. Probabilities can change in more ways than could be implicated by single samples. Moreover, even if multiple samples are available, storing previous information in a running estimate of a model and incrementally updating that model to the nearest best-fitting model has a number of other disadvantages. There are issues with storing information as the current estimate and there are issues with only incrementally updating a single working model. These disadvantages are developed below and hypothesis testing is offered as an alternative account.

Information Contained in Multiple Samples

The simplest critique of trial-by-trial updating is that single samples cannot reliably be used to determine the trajectory of hypothesis updating. Any single sample may be ambiguous in implication. For example, consider the case where a distribution is assumed to be bimodal. Any particular event has two possible interpretations; either one or the other mean has shifted. A delta-model updating algorithm will not know which parameters to change to make the recent sample more likely if those parameters are correlated.

This calls into question trial-by-trial updating if trial-by-trial is taken to mean that every sample is used independently to immediately generate a better estimate of the probability. However, it should be noted that this does not rule out delta-rule updating. If the model is complex enough to require converging information from multiple samples, it should be noted that the delta updating model can be amended to operate on multiple samples. If single samples underdetermine the next best fitting model, delta-rule mental model updating may rely on a window over multiple samples. However, the limitation of the delta-rule model is not solely due to its trial-by-trial implementation. There is information that can be extracted from multiple samples which cannot be detected by a delta-rule model.

The delta-rule model assumes the form of the frequency distribution it is using to

represent events. When new data arrives, this distribution is updated to make those events more likely. The delta-rule model assumes a class of models and uses its constraints to improve the current working estimate of the probability. However, it is possible for the form of the stochastic process to itself change.

As an example, consider the case where a unimodal distribution changes to a bimodal distribution. If a unimodal distribution is assumed, the delta-rule model would become more volatile and would be pulled sporadically between two means. This does not occur in human estimates. Humans are quite capable of tracking changes in distributions which require multiple samples to detect. For instance, they can notice a change from a unimodal to a bimodal distribution [31] or a narrow distribution to a wide distribution [32].

In general, humans can infer the broader classes of models which their particular probability estimates are instances of. They can learn the general form of a probability distribution and the rules that govern how the distribution changes [29, 33, 34, 35].

As discussed, the delta-rule updating model cannot learn these sorts of constraints on models. Constraints on models are not themselves not contained in the model but rather determine the form of models and determine how information is used to update those models. Constraint learning thus operates outside of delta-model learning. No incremental updating is possible if what is learned is rules for updating. Rules for updating here refers to how the class of possible probability models is constrained such that the future evidence is used to find a better fitting model within a class. When different classes of models account for the same data, here the importance of hypothesis testing becomes evident. When models in two different classes account for recent samples differently, some method must be used to determine which model is preferable.

One method of switching between classes of models or adopting different constraints on postulated distributions is hierarchical Bayesian inference. The hierarchy in hierarchical Bayesian inference refers to the structure of the hypothesis space that classes of models exist within. Moving a level up in hierarchical hypothesis space means making a model less complex or less specific. Moving a level down in hypothesis space means constraining the model, making it more specified. An example of a model increasing in complexity is adding a predictor to a regression equation. The model gets more complex as its estimates of a variable depend on context, but model predictions become more precise. The more complex model produces better estimates for the expected values of the dependant variable.

Hierarchical Bayesian inference offers a method to distinguish which model ought to be adopted to explain recent evidence (see [29] or [34] for more detail). Increasing model complexity tends to increase model fit. Samples are better explained by the increased explanatory power of increasingly specified models. However, there is a trade-off between

model-fit to previous observations and model generalizability to future data; this is the concern with overfitting models.

The real ingenuity of the hierarchical Bayesian approach is in the solution it offers to overfitting. Hierarchical Bayesian reasoning naturally favours simpler models. This has been referred to as the size principle [34] and the minimum principle [36]. Gallistel et al. [4] describe the principle as a built-in Occam's razor. In hierarchical Bayesian reasoning, a complex model with more parameters is given the same prior likelihood as a simpler model with fewer parameters. Complex models distribute their probability over a larger space and marginalize over more sub-optimal parameterizations. The increase in fit offered by the intersections of parameters is counteracted by the smaller weight assigned to the best fits and the averaging of fit over a broader hypothesis space. This creates a bias against more complex models.

To ground the broad discussion back in the current research, it is worthwhile to explore how the Bernoulli process used in this task itself exists implicitly in a hierarchical model space. The single parameter in the Bernoulli process represents a constraint on a higher-order class of binary stochastic processes. In a Bernoulli process, samples are independent and identically distributed (IID). That means that, given the Bernoulli process, each dot color is independent of the last dot colour. However, this is a special case of a broader class of possible transition probabilities. It is possible that both dot colors have different probabilities of being followed by either dot color. In a Bernoulli process, those probabilities happen to be shared over the previous states.

The two transition probabilities being uniform across previous states is a constraint on the broader class of models which a Bernoulli process is a subset of. But people are capable of learning either individual transition probabilities [37] or collapsed transition probabilities (as in this Bernoulli task). This suggests that they can entertain multiple explanations for the same data and apply different constraints at different times. This suggests that humans can move around hierarchical hypothesis space.

Of note, transition probabilities in such an increasingly complex Bernoulli task have been used to argue that hierarchical reasoning is necessary part of tracking nonstationary processes. Heilbron and Meyneil [30] argue that previous tasks have been too simple to truly distinguish flat from hierarchical reasoning. The experimental design used to argue this point was similar to previous experimental work using Bernoulli processes, however, two transition probabilities were used rather than was the one probability assumed to be uniform across previous draws. The probabilities changed together at distinct change-points. This created a dependency between the two transition probabilities; if a change was detected in one transition probability, a change in the other transition probability could be

inferred. Only the hierarchical inference algorithm could be sensitive to this dependency. Human estimates, like only the hierarchical learning model, were sensitive to these parallel change-points.

It has thus far been argued that relying on single samples to update estimates is too severe a restriction on the set of possible inferences that can be obtained from sample information. Trial-by-trial updating is limited partly because single samples are not enough to determine the direction a model needs to change in. Statistical information obtained from multiple coincident samples may go so far as to implicate shifts in the constraints on the probability model currently being used to determine the probability of events. This kind of model updating is not possible given a flat delta-rule updating model. Rather, model comparison seems to be required.

Information Stored in Working Estimate

This next section further problematizes the delta-updating model by arguing that storing information in a current estimate of the probability and incrementally updating this estimate is insufficient to account for how estimates must be updated by humans. Some hypotheses cannot be moved to via incremental updating of a single working estimation in an assumed parameter space. Information is lost when only the current estimate of a working mental model is used. If only one working estimate of probabilities is remembered the explanatory power of completely disparate hypotheses may be lost and unrecoverable.

Previously, constraint learning was only discussed in the context of coincident samples being suspiciously well explained by a particular model. This seems to imply that learning constraints also implies a best-fit model. This is not so. Constraints can be on the broader hypothesis space without themselves implicating one best-fit explanatory model or even a particular class of models. Information may ambiguously support a number of distinct hypotheses or disparate models.

In psychology there exists a term which is used to describe the case where evidence ambiguously supports a number of discrete hypotheses: confounds. Information contained in samples is often confounded. Evidence for more than one explanation can be explained away by an alternative hypothesis when new evidence is gathered.

Humans are quite capable of learning constraints on models through ambiguous information. For example, Griffiths and Tenenbaum [38] had participants rate the likelihood that objects were 'blickets' based on the output of a 'blicket detector'. In the indirect screening-off condition, participants watched two objects go into the detector. The blicket

detector went off, signaling the presence of at least one blicket. This information ambiguously supports three potential hypotheses. Either one or the other or both objects going into the machine are blickets. In the subsequent trial, participants watched one object go into the blicket detector and the blicket detector did not go off. They correctly inferred that the other object that did not go through the machine was likely to have been the blicket.

No estimate of the probability of each item being a blicket supported this conditional inference that the absence of a blicket signal from one object implicated the blicket status of the other object. Rather, the information gained in the previous trials determined constraints on the set of probable worlds which could be inferred given further information. The meaning of the previous information did not determine a single mental model which recursively stored previous information from the time series. Rather, it determined the way in which information about one state of affairs related to another state of affairs.

Explanatory models accounting for the same evidence equally well can differ from one another quite dramatically. There may be no smooth incremental improvement between models over the parameter space if only an intersection of parameters determines that the model fits well. This will be the case if it is particular intersections of states of affairs that best account for data, rather than there being a monotonic increase in model fit given any incremental shift in parameters. However, humans are quite capable of switching between two explanatory models which account for the data very differently. This ability is apparent, for example, in bistable perception [39].

To return the discussion to the nonstationary doubly stochastic processes, a neurologically plausible account of multiple hypothesis comparison (rather than single estimate incremental updating) has been offered for time-series data. This is particle sampling [40, 7, 10]. Particle sampling is a Monte Carlo approximation of optimal Bayesian inference for time-series data. Rather than there being a single estimate of the probability which is updated according to incoming evidence, multiple simultaneously entertained hypotheses are represented by a number of particles. The proportion of particles currently entertained represents the strength of each hypothesis. The number of particles per hypothesis is what is updated with incoming sample information. The distribution of particles approximates the posterior distribution over the hypothesis space of explanatory models [10]. Of particular interest, Prat-Carrabin et al. [7] fit a number of different models to participant estimates of a doubly stochastic stepwise nonstationary process. The models include a particle filter model, but also many of the other models proposed by various authors discussed above [8, 21, 6]. They found that a particle filter model best described participant estimates.

That Which Constitutes a Sample

Finally, it is also worth noting that identifying units that count as samples can itself be learned through probabilistic inference. Single samples are necessary to update a trial-by-trial model to make those samples more probable, but sometimes the segmentation of information into discrete units is itself probabilistic inference. In statistical learning, transition probabilities are learned and used to segment continuous streams into categorical units. This method of inference was first proposed to explain parsing acoustic signal into language. Rather than there being distinct pauses in speech between words as one might expect, word boundaries are more consistently marked by the low probability of the phonological transitions between words in sound sequences [41].

Within a perceptual unit, events are highly predictable, and it is the low transition probabilities at the end of the unit which signals its end [42]. Hard, Recchia and Tversky [43] use the same principle to explain how it is we segment an unfolding event into sub-procedural actions. The ends of actions are also marked by low transition probabilities between action subroutines. If a person is engaging in an activity, for example, doing laundry, their actions are highly predictable. However there is increased entropy at the boarder of actions where it becomes more variable what a person engages in next.

Summary

Even while the delta-rule model offers a number of advantages for simple model updating, there must also be a more sophisticated mechanism detecting more complex changes in probabilities over time. As a general principle for updating estimates of probabilities as they change over time, the delta-rule fails to explain all possible probabilistic inferences. The principle of hypothesis testing as used in change-point detection algorithms has much broader applications. Hypothesis testing is implied by more general and sophisticated models which can detect highly complex changes in probabilities over time. While optimal hierarchical Bayesian inference may be computationally intractable, its approximations such as particle filtering [10] offer a neurologically plausible account of human estimation [40] and they have been found to fit human estimates better even than delta-updating models [7].

4.5 Conclusion

In our experiment, effort was not sufficient to explain response maintenance. Despite an increase in motor cost, some participants nonetheless maintained their estimates over long series of trials. They did not use each sample to update their estimates of the probability. Response maintenance, despite increased difficulty, emerged naturally in some participant estimates of the probability. But this strategy is not universal. Individual variation was large in both groups. In the Automatic participant data, a large range of updating frequencies is particularly evident. This supports there being a larger range of updating strategies. It is certainly possible, given a simple learning task, to use each piece of evidence to update a mental model. However, the possibility of altering models to incorporate incoming evidence and improve a working model should not be over-generalized.

Our data suggest response maintenance is not an artifact of previous designs but points to a genuine strategy for probability estimation. This is consistent with previous literature finding improved model fitting given two-step probability estimation procedures for which a threshold prevents immediate updating. This is also consistent with previous literature which finds that human beings are sensitive to probability structures which are not determined by single samples and only become likely when sufficient evidence accumulates.

References

[1] Patricia K Kuhl. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics*, 50(2):93–107, 1991.

[2] Naomi H Feldman, Thomas L Griffiths, and James L Morgan. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4):752, 2009.

[3] Nathaniel Delaney-Busch, Emily Morgan, Ellen Lau, and Gina R Kuperberg. Neural evidence for Bayesian trial-by-trial adaptation on the n400 during semantic priming. *Cognition*, 187:10–20, 2019.

[4] Charles R Gallistel, Monika Krishan, Ye Liu, Reilly Miller, and Peter E Latham. The perception of probability. *Psychological Review*, 121(1):96, 2014.

[5] Matthew Ricci and Randy Gallistel. Accurate step-hold tracking of smoothly varying periodic and aperiodic probability. *Attention, Perception, & Psychophysics*, 79(5):1480–1494, 2017.

[6] Mel Win Khaw, Luminita Stevens, and Michael Woodford. Discrete adjustment to a changing environment: Experimental evidence. *Journal of Monetary Economics*, 91:88–103, 2017.

[7] Arthur Prat-Carrabin, Robert C Wilson, Jonathan D Cohen, and Rava Azeredo da Silveira. Human inference in changing environments with temporal structure. *Psychological Review*, 2021.

[8] Matthew R Nassar, Robert C Wilson, Benjamin Heasly, and Joshua I Gold. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37):12366–12378, 2010.

[9] Matthew R Nassar, Katherine M Rumsey, Robert C Wilson, Kinjan Parikh, Benjamin Heasly, and Joshua I Gold. Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, 15(7):1040–1046, 2012.

[10] Scott D Brown and Mark Steyvers. Detecting and predicting changes. *Cognitive psychology*, 58(1):49–67, 2009.

[11] Gordon H Robinson. Continuous estimation of a time-varying probability. *Ergonomics*, 7(1):7–21, 1964.

[12] Fuat Balci, David Freestone, and Charles R Gallistel. Risk assessment in man and mouse. *Proceedings of the National Academy of Sciences*, 106(7):2459–2463, 2009.

[13] Fuat Balci, David Freestone, Patrick Simen, Laura Desouza, Jonathan D Cohen, and Philip Holmes. Optimal temporal risk assessment. *Frontiers in Integrative Neuroscience*, 5:56, 2011.

[14] Mark Steyvers and Scott Brown. Prediction and change detection. *Advances in neural information processing systems*, 18, 2005.

[15] Tadeg Quillien and Chris Lucas. The logic of guesses: how people communicate probabilistic information. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

[16] Lea K Krugel, Guido Biele, Peter NC Mohr, Shu-Chen Li, and Hauke R Heekeren. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences*, 106(42):17951–17956, 2009.

[17] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[18] Timothy EJ Behrens, Mark W Woolrich, Mark E Walton, and Matthew FS Rushworth. Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9):1214–1221, 2007.

[19] Jill X O'Reilly, Urs Schüffelgen, Steven F Cuell, Timothy EJ Behrens, Rogier B Mars, and Matthew FS Rushworth. Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, 110(38):E3660–E3669, 2013.

[20] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[21] Robert C Wilson, Matthew R Nassar, and Joshua I Gold. Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476, 2010.

[22] Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[23] Herbert Hoijtink, Joris Mulder, Caspar van Lissa, and Xin Gu. A tutorial on testing hypotheses using the Bayes factor. *Psychological methods*, 24(5):539, 2019.

[24] Mattias Forsgren, Peter Juslin, and Ronald van den Berg. Further perceptions of probability: in defence of trial-by-trial updating models. *BioRxiv*, 2020.

[25] Joshua R De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12, 2015.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[27] Florent Meyniel, Daniel Schlunegger, and Stanislas Dehaene. The sense of confidence during probabilistic learning: A normative account. *PLoS computational biology*, 11(6):e1004305, 2015.

[28] Ward Edwards, Harold Lindman, and Lawrence D Phillips. Emerging technologies for making decisions. 1965.

[29] Thomas L Griffiths and Joshua B Tenenbaum. From mere coincidences to meaningful discoveries. *Cognition*, 103(2):180–226, 2007.

[30] Micha Heilbron and Florent Meyniel. Confidence resets reveal hierarchical adaptive learning in humans. *PLoS computational biology*, 15(4):e1006972, 2019.

[31] Peter A. V. DiBerardino, Alexandre L. S. Filipowicz, James Danckert, and Britt Anderson. Plinko: Eliciting beliefs to build better models of statistical learning and mental model updating, 2021.

[32] Hanbin Go, James Danckert, and Britt Anderson. Saccadic eye movement metrics reflect surprise and mental model updating. *Attention, Perception, & Psychophysics*, pages 1–13, 2022.

[33] Leah Henderson, Noah D Goodman, Joshua B Tenenbaum, and James F Woodward. The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. *Philosophy of Science*, 77(2):172–200, 2010.

[34] Amy Perfors, Joshua B Tenenbaum, Thomas L Griffiths, and Fei Xu. A tutorial introduction to Bayesian models of cognitive development. *Cognition*, 120(3):302–321, 2011.

[35] Joshua B Tenenbaum, Thomas L Griffiths, and Sourabh Niyogi. Intuitive theories as grammars for causal inference. *Causal learning: Psychology, philosophy, and computation*, pages 301–322, 2007.

[36] Jacob Feldman. Bayes and the simplicity principle in perception. *Psychological review*, 116(4):875, 2009.

[37] Raymond Y Cho, Leigh E Nystrom, Eric T Brown, Andrew D Jones, Todd S Braver, Philip J Holmes, and Jonathan D Cohen. Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. *Cognitive, Affective, & Behavioral Neuroscience*, 2(4):283–299, 2002.

[38] Thomas L Griffiths and Joshua B Tenenbaum. Theory-based causal induction. *Psychological review*, 116(4):661, 2009.

[39] Tai Sing Lee and David Mumford. Hierarchical Bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

[40] Lei Shi, Thomas L Griffiths, Naomi H Feldman, and Adam N Sanborn. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic bulletin & review*, 17(4):443–464, 2010.

[41] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

[42] Niels Chr Hansen, Haley E Kragness, Peter Vuust, Laurel Trainor, and Marcus T Pearce. Predictive uncertainty underlies auditory boundary perception. *Psychological science*, 32(9):1416–1425, 2021.

[43] Bridgette Martin Hard, Gabriel Recchia, and Barbara Tversky. The shape of action. *Journal of experimental psychology: General*, 140(4):586, 2011.

[44] Inbal Arnon and Neal Snider. More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82, 2010.

[45] Gina R Kuperberg and T Florian Jaeger. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59, 2016.

[46] Elyse H Norton, Luigi Acerbi, Wei Ji Ma, and Michael S Landy. Human online adaptation to changes in prior probability. *PLoS computational biology*, 15(7):e1006681, 2019.

[47] Amy Perfors, Joshua B Tenenbaum, and Terry Regier. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338, 2011.

[48] Amy Perfors, Joshua B Tenenbaum, and Terry Regier. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338, 2011.

[49] Fei Xu and Joshua B Tenenbaum. Word learning as Bayesian inference. *Psychological review*, 114(2):245, 2007.