UNIVERSITY OF CALIFORNIA,
IRVINE

A Mathematical Explication of Human Psychology

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Gregory E. Alexander

Dissertation Committee:
Professor William H. Batchelder, Chair
Professor Michael Lee
Professor Louis Narens

2017

DEDICATION

To Rory V. Alexander

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank the many people in my life that made this all possible. I would also like to thank the National Science Foundation Grant #1534471 awarded to William H. Batchelder, PI.

CURRICULUM VITAE

Gregory E. Alexander

EDUCATION

Doctor of Philosophy in Psychology **2017**
University of California, Irvine *Irvine, California*

Masters of Arts in Cognitive Science **2014**
University of California, Irvine *Irvine, California*

Bachelor of Arts in Cognitive Science **2011**
University of California, Irvine *Irvine, California*

PROFESSIONAL AFFILIATION

Society for Mathematical Psychology. **2011–2017**

FELLOWSHIPS AND AWARDS

UCI Cognitive Science Department Summer Research Support **2015**

UCI Advisor Summer Research Support **2014**
Institute of Mathematical Behavioral Sciences Research Support **2013**
UCI Student Travel Award **2013**
Undergraduate Research Opportunities Program Grant **2010**

RESEARCH EXPERIENCE

Research Assistant IV **2010–2011**
Department of Cognitive Science, UCI *Irvine, California*

Research Assistant **2009–2011**
Department of Neurobiology of Learning and Memory UCI *Irvine, California*

Independent Student Researcher **2009–2010**
Department of Cognitive Science, UCI *Irvine, California*

INVITED LECTURES

Psychology 229: Stochastic Modeling **2015**
University of California, Irvine *Irvine, California*

Psychology 234A: Mathematical Models **2014**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Psychology 120H: History of Psychology **2015**
University of California, Irvine *Irvine, California*

Psychology 140C: Cognitive Science. **2015**
University of California, Irvine *Irvine, California*

Psychology 140M: Human Memory **2013**
University of California, Irvine *Irvine, California*

Psychology 143P: Human Problem Solving **2012**
University of California, Irvine *Irvine, California*

Psychology 9B: Introduction to Psychology **2012**
University of California, Irvine *Irvine, California*

Psychology 160A: Abnormal Psychology **2012**
University of California, Irvine *Irvine, California*

INDEPENDENT CONSULTANT

Statistical consultant **2013-2015**
Medical Care Corporation *Newport Beach, California*

Consultant **2012**
Office for Research Information System (ORIS)

REFEREED JOURNAL PUBLICATIONS

Alexander, G. E., Satalich, T. A., Shankle, W. R., & Batchelder, W. H. (2015). A Cognitive Psychometric Model for the Psychodiagnostic Assessment of Memory Related Deficits. Psychological Assessment.

Batchelder, W. H. & Alexander, G. E. (2013). Discrete-state Models: Comment on Pazzaglia, Dube, and Rotello (2013). Psychological Bulletin.

Batchelder, W. H., & Alexander, G. E. (2012). Classical Insight Problem Solving: A Critical Examination of the Possibility of Psychological Theory. Human Problem Solving.

TECHNICAL REPORTS

Alexander, G. E. (2015). Technical Report: An Analysis of AVLT study using a Hidden Markov Model.

Alexander, G. E., & Batchelder, W. H. (2013). Technical Report: A Hidden Markov Model for Psychodymamic Assessment of Memory.

ABSTRACTS, PAPER, AND POSTER PRESENTATIONS

Alexander, G. E. (2016). Deconstructing the Transient Nature of Cultural Truth. Mathematical Psychology Conference. (Rutgers, New Jersey)

Alexander, G. E. (2016). No More Know-it-alls. 54rd Edwards Bayesian Research Conference.

Alexander G. E., Batchelder W. H., Shankle W. R. Measuring Cognitive Processes Affected by Alzheimers Disease Using Markov Models. Clinical Trials on Alzheimer's Disease CTAD 2015. Oral Presentation. Barcelona, November 2014 (accepted).

Alexander, G. E. (2015). Measuring Cognitive Variables Using a Hidden Markov Model. Mathematical Psychology Conference. (Newport Beach, California)

Alexander, G. E. (2015). Alzheimers disease isnt all bad. Its just mostly bad. 53rd Edwards Bayesian Research Conference.

Alexander G. E., Satalich T. A., Batchelder W. H., & Shankle W. R. (2014). Application of Markov Models to Word Recall Tasks to Improve Cognitive Assessment Capability for Better Trial Management. CTAD 2014. Oral Presentation. Philadelphia, November 2014.

Alexander, G. E. (2014). Finding Hidden Variables using a Hidden Markov Model. Colloquium Series. University of California, Irvine.

Alexander, G. E., & Batchelder, W. H. (2014). The Effects of Loosing Your Marbles; A Tale by Free Recall. 52th Edwards Bayesian Research Conference

Satalich, T. A., Alexander, G. E., Batchelder, W. H., & Shankle, W. R. (2014). Markov Models Detect Vitamin E and Donepezil Treatment Effects in ADCS MCI Trial. Alzheimers Association International Conference.

Alexander, G. E., Satalich, T., Batchelder, W. H., & Shankle, W. R. (2014). Markov Models of ADAS-Cog Memory Task Greatly Improve Signal-to-Noise Ratio For Detecting Change. Alzheimers Association International Conference.

Alexander, G. E. (2013). Stimulus Similarity in Continuous Recognition Memory. Mathematical Psychology Conference. (Postdam, Germany)

Alexander, G. E. (2012). A Statistical Development and Comparison of Two Useful Recog-

nition Memory Models. Mathematical Psychology Conference. (Ohio)

Batchelder, W. H., & Alexander, G. E. (2012). Doubling down on Double High Threshold. 50th Edwards Bayesian Research Conference

Alexander, G. E. (2011). Workshop MPT modeling. Mathematical Psychology Conference, Boston

Alexander, G. E., Rabideau, C. M., & Phoong, G. W. (2010). Gist and Item Memory: The Effects of Multiple Exposures on Memory. Poster presented at the Tenth Annual Stanford Undergraduate Psychology Conference, Stanford, CA.

Alexander, G. E., Rabideau, C. M., & Phoong, G. W. (2010). The Effects on Memory after Multiple Presentations. Poster presentation at the University of California, Irvine Undergraduate Research Symposium, Irvine, CA.

PATENT

Alexander, G.E., and Shankle, W.R. 2015. Assessing cognition using item-recall trials with accounting for item position. U.S. Patent Application PCT/US2015/015282, filed February 2015. Patent Pending.

ABSTRACT OF THE DISSERTATION

A Mathematical Explication of Human Psychology

By

Gregory E. Alexander

Doctor of Philosophy in Psychology

University of California, Irvine, 2017

Professor William H. Batchelder, Chair

Our scientific knowledge of human behavior has taken great leaps with the formalization of quantitative psychology. This dissertation is an amalgamation of mathematical models in the field of psychology, specifically as it pertains to higher order cognition. The goal is to provide a variety of useful contributions to psychology in three unique areas of the field.

The first focuses on Signal Processing Models in recognition memory. I begin by outlining the two most popular models and describing their mathematical properties. This is done to promote both models usefulness as measurement tools, regardless of their mathematical differences. I continue by developing a novel extension for each model to further elucidate their usefulness in psychology.

The second area of research discussed in this dissertation moves away from purely theoretical applications of mathematical models towards real-world applications of a stochastic system. Here, we develop and explore a Hidden Markov Model for memory deficits, with the goal of understanding dementia. Since clinical trials contain a variety of memory tests, a second paper devoted to further understanding memory decay in Alzheimers is provided.

Finally, the last two chapters of this dissertation focus on decision making as it applies to information pooling techniques. We utilize the mathematical concepts developed throughout

the dissertation in order to identify an area of improvement for models in current use, and offer an innovative new interpretation of existing theory. The final paper explores a natural extension to the theory for continuous-type responses, and outlines further opportunities for additional research in this area.

Chapter 1

Likelihood Analysis of the Signal Detection and Double High Threshold

In the study of recognition memory, two prominent mathematical theories that have thus far stood the test of time have been Signal Detection theory and Threshold theory. Recent work on these models has demonstrated that the basic versions of each model are statistically equivalent and has called for other scientifically motivated methods to differentiate the two. The focus of this paper is to demonstrate that differences between the models can be revealed from a mathematical approach. In this spirit, we evaluate the likelihood functions of each model and the corresponding Fisher Information.

1.1 Introduction

In this paper the likelihood functions of two popular models for Yes/No recognition memory experiments are compared. The models are the Signal Detection theory (SDT) model with Gaussian familiarity distributions (Mcmillan, 2004), and the Double High-Threshold (2HT) model (Snodgrass, 1988) that is a member of the class Multinomial Processing Tree (MPT) models. Each parametric model provides an account of select complex human memory processes believed to be involved in recognition memory. With the ever growing literature on these two models researchers have opined as to their preferred theory regardless of the fact that the exact tally of latent processes involved in human memory is unknown and quite possibly unknowable.

Naturally as the number of papers dedicated to proving one model superior to the other increases, so do the model selection techniques. Within this ever growing literature, mathematical techniques have evolved from simply a goodness of fit measure (usually denoted by minus twice the log likelihood), to techniques such as the Akaiki Information Criterion (AIC), Bayesian Information Criterion (BIC), Normalized Maximum Likelihood (NML), and finally Bayes Factors which account for model flexibility. For example, AIC accounts for model flexibility by the number of parameters a model has, whereas BIC penalizes a model by the sample size. Since neither account for the functional form of the model, researchers have turned to NML which denotes model flexibility as all possible data patterns a specific model can account for. While all of these techniques have demonstrated support for a particular model on selected experimental results no distinct winner has emerged. Instead, this vacillation of methodology has led the scientific community to be torn between what we consider equally suitable explanations in light of obvious limitations.

It is possible that neither model emerges as a clear winner despite these more complex model selection techniques because these two models are more alike than previously understood.

For instance, the basic SDT and 2HT models each have two free parameters, so measures such as the AIC will not be able to differentiate the models, and since both are tested on the same sample BIC may not help either. Furthermore, the two models are statistically equivalent (Alexander & Batchelder, 2013) so both entail the same probability distributions, and thus NML will not show a difference.

For techniques aimed at testing the different theories, the likelihood function stands as a pivotal measure. This dependency on the likelihood function warrants a closer look at the relationship each model has to a normative set of probability distributions. More specifically, a particular model's likelihood function will have no solution on any set of observable values outside the predictive region, and thus the model will inherently fall short of fitting the data by any measure. Therefore, this paper will focus on the models' likelihood functions and their implicit solutions. For example, since the two basic forms of each theory are statistically indistinguishable, a more mathematical approach may elucidate differences.

Our goal is not to argue that one of these models is better than the other, but instead to compare and contrast the two models. In this comparison, we first consider these models on purely statistical grounds. We find that the models are statistically equivalent, namely in that they both entail exactly the same set of probability distributions over the sample space for the Yes/No recognition memory experiment.

We will begin by defining the models and their prediction space, followed by a study of their nth order partial derivative, from which the maximum likelihood estimators are found. Next, the amount of information extracted from the random variables for each model is calculated using Fisher's Information. Two useful results emerge from the Fisher's Information: 1) The inverse Fisher's Information details the precision for which a parameter can be estimated, 2) we obtain the necessary number of samples for an acceptable degree of error. Finally, issues of numerical instability are explored through a parameter sensitivity test.

1.2 Signal Detection Model

The one-dimensional SDT model (e.g. Macmillan & Creelman, 2005) for recognition memory assumes a recognition judgment is derived from the familiarity strength of an item on a continuum of memory states. Familiarity strength is conceptualized as a continuous random variable on a familiarity axis. For each class of items, there is a probability distribution over the familiarity axis. While all items have some degree of familiarity, the role of the study phase is to increase familiarity with a predetermined set of target items. Generally, the amount of overlap between the distributions of two classes of items, (e.g. old/new), indicates the discriminability level between them, and extra experimental manipulations such as word repetition or word saliency can have an effect on the amount of overlap between the distribution of the two classes of items.

Within this framework it is not possible to know exactly whether an item is old or new, so SDT postulates that a decision maker (DM) is only aware of an item's familiarity strength relative to an established criterion. Thus in order to decide the class membership of the item, a threshold, τ, is predetermined[1] by the DM. When an item's familiarity strength is above the pre-established threshold on the familiarity axis, the DM responds in favor of the strengthened class of items. In an old/new recognition memory task, a familiarity strength above τ is considered old and thus a DM responds 'old' and items with familiarity strengths that fail to surpass the threshold are judged 'new'.

To this date, the most common distributional assumption of the latent familiarity index for each class of items is a Gaussian distribution. The use of a Gaussian distribution greatly simplifies the problem since the distribution is completely characterized by two parameters. We retain this distributional assumption and allow $\{\mu_o\ ,\ \sigma_o\}$, and, $\{\mu_n\ ,\ \sigma_n\}$, to be the parameters for the old and new class of items, respectively. Since the role of the study phase

[1]While it is generally accepted that DMs utilize a threshold to determine an item's membership, there seems to be no mention about the construction of this threshold.

is to increase the familiarity levels of old items, we expect that on average the old class of items will elicit a greater degree of familiarity expressed as $\mu_o \geq \mu_n$. Of course, the standard deviations are measures of the variability of the class of items so they are $\sigma_o > 0$ and $\sigma_n > 0$.

As it stands, the model exemplifies the theoretical assumptions of SDT thus permitting one to draw conclusions of psychological variables using the parameters. Now the difficulty of solving for the parameters rests on methodological limitations concerned with the availability of sufficient information. With only two degrees of freedom in an old/new recognition experiment the model parameters cannot be uniquely identified. To avoid the identifiability problem caused when a model has more parameters than degrees of freedom (d.f.) (Bamber, D., & van Santen, J. P. H. 2000), we reduce the number of free parameters to only two by fixing the values of prespecified parameters. Without loss of generality, $\mu_o = d'/2$, $\mu_n =$ -$d'/2$, where $d' \in [0,\infty)$, $\sigma_o = 1$, and $\sigma_N = 1$ and $\tau \in (-\infty, \infty)$. A pictorial account of this can be found in Figure 1.

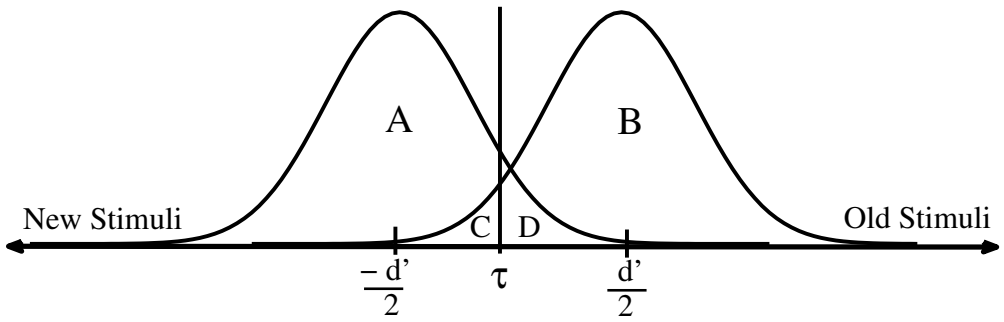

Figure 1.1: Equal-Variance Signal Detection Model; A- Correct Rejection; B- Hit Rate; C- Miss; D- False Alarm

The two remaining free-to-vary parameters are d' and τ, where parameter d' is known as the discrimination index and indicates the strength of the signal distribution in contrast

to the noise distribution, and τ represents the pre-established threshold mentioned above. Under strictly controlled experimental conditions, the simplified model is sufficient for an experimenter to identify the model's latent variables in an old/new recognition memory experiment. However, the requirement of strict experimental conditions is not always met, so it has been advocated that the model be modified to allow unequal variance Gaussian distributions (Wixted 2007a).

1.3 Basic High Threshold Model

Unlike the continuous nature of SDT, 2HT assumes that a recognition judgment is based on an item's membership to one of three discrete memory states achieved by two high thresholds. The three states are Detect, Discriminate, and Guessing states, which correspond to different levels of information classes. The mutually exclusive states form the basis of all judgments by fully describing the state of memory for a test item. The role of the study trial is to facilitate a transition away from a guessing state characterized by a complete lack[2] of sensory information pertaining to the item. Within this guessing state, items are judged based on a probabilistic process denoted by γ, where $\gamma \in [0, 1]$. Since no pertinent information is assumed to exist within this state, a response is often viewed as a outcome from a purely guessing process.

Similarly to SDT, threshold models assume that during a learning trial, items are reinforced, but unlike SDT, the item is posited to enter a decisive state where knowledge of an item's true classification is known. The occurrence of such an event happens by exceeding one of two thresholds captured by the probabilities D_o and D_n, for old and new items, respectively, where $D_o \in [0, 1]$ and $D_n \in [0, 1]$. Once an item is situated in either the detect or discrimi-

[2]In a old/new recognition memory task episodic memory is studied, thus a lack of sensory information is meant as a mnemonic device to disambiguate items that hold no meaning to the task rather than to describe novel stimuli.

nate state a judgment on the test will always correspond to the state's identity. Naturally, the model predicts a DM is capable of knowing that an item belongs to the old (studied) or new (unstudied) set of items. In fact, it is easy to see that if an experiment involves the presentation of a single item on the study phase and is tested against a single distractor on the test, the DM would certainly 'know' which item was old, presumably without the need of forming a decision criterion. While the gedankenexperiment may be too simplistic, it reveals a fundamental unexplored issue of when exactly a DM becomes incapable of knowing an item's true assignment and is therefore forced to rely on a generated decision criterion as predicted by SDT.

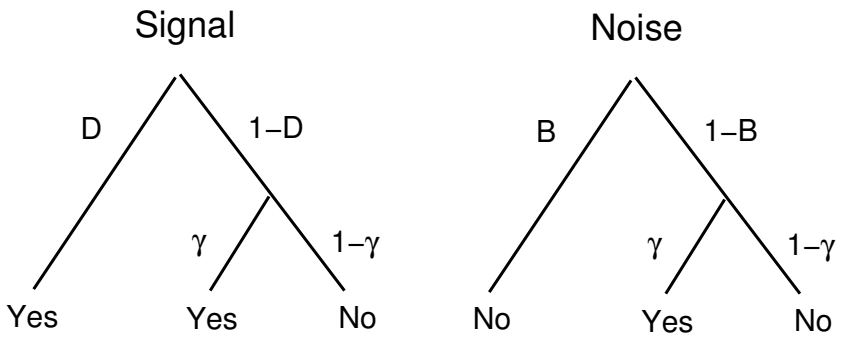

Figure 1.2: Basic Double High Threshold Model. Tree structure of 2HT model with two parameters. Paths down the tree diagram of figure 2 show the possible ways of producing an 'old' or 'new' response for the item class.

Once again, methodological difficulties create limitations when attempting to solve the precise and well-formulated mathematical problem of estimating parameter values of the model. Thus the 2HT, just as the SDT model, has to make some limiting assumptions in order to reduce the three free varying parameters down to two. Researchers normally accomplish

this by equating D_o and D_n. Support for collapsing the two free parameters to one comes from the mirror effect in recognition memory (Adams, 1985). The mirror effect reflects the phenomena often found in recognition memory experiments, that when the hit rate (H) increases, the false alarm rate (FA) decreases, and when H decreases FA increases. Although the reduction in free parameters is a necessary step for model identifiability, it poses problems of its own. By coalescing the two parameters, we are forced to assume that the probability of detecting an old item, which is dependent on the amount of information stored during the study trial, is the same as the probability of discriminating a new item. Proponents of the 2HT model have therefore sought methods that would allow for D_o and D_n to differ.

1.4 Mathematical Comparison

Let X and Y be two independent Bernoulli random variables with success probabilities $\{p, q\}$. Denote the marginal probability distributions of a single Bernoulli event x_1 and y_1 as $f_X(x_1) = p^{x_1}(1-p)^{(1-x_1)}$ and $f_Y(y_1) = q^{y_1}(1-q)^{(1-y_1)}$ and the joint probability of M_1 and M_2 independent Bernoulli trials $X_1, \ldots, X_{M_1}$ and $Y_1, \ldots, Y_{M_2}$ as

$$f_X(x_1, \ldots, x_{M_1}) = \prod^{M_1} p^{x_i}(1-p)^{(1-x_i)} = p^{\bar{x}}(1-p)^{(M_1-\bar{x})}$$

and

$$f_Y(y_1, \ldots, y_{M_2}) = \prod^{M_2} q^{y_j}(1-q)^{(1-y_j)} = q^{\bar{y}}(1-q)^{(M_2-\bar{y})}$$

where $\bar{x} = \sum x_i$ and $\bar{y} = \sum y_j$.

Parametric models, such as SDT and 2HT, redefine the parameters $\{p, q\}$ of the marginal probability distribution for each random variable X and Y. In particular, psychological models map latent properties (i.e. parameters) to behavior or outcomes (i.e. instances of

the random variables X and Y). Unable to directly observe the latent properties of interest, we rely on statistical methods such as maximum likelihood to explore the data generating process. We begin by defining the parametric models and analyzing the likelihood functions for a comprehensive analysis of the models.

Definition 1.a Let X, Y be discrete independent bivariate random vectors, with X representing the scored response for an old item, and Y representing the scored response for a new item, and let the likelihood function $L(\Theta_{SDT}|X, Y)$ be defined on the cross product $M_1 \times M_2 = \{(\bar{x}, \bar{y}) : \bar{x} \in M_1, \bar{y} \in M_2\}$, as:

$$L(\Theta_{SDT}|X,Y) = \begin{cases} \prod_{i}^{M_1} f_{\Theta_{SDT}}(x_i) = \Phi(\frac{d'}{2} - \tau)^{\bar{x}} \Phi(-\frac{d'}{2} + \tau)^{M_1 - \bar{x}} & (1) \\ \prod_{j}^{M_2} f_{\Theta_{SDT}}(y_j) = \Phi(-\frac{d'}{2} - \tau)^{\bar{y}} \Phi(\frac{d'}{2} + \tau)^{M_2 - \bar{y}} & (2) \end{cases}$$

Definition 1.b Let X, Y be discrete independent bivariate random vectors, with X representing the scored response for an old item, and Y representing the scored response for a new item, and let the likelihood function $L(\Theta_{2HT}|X, Y)$ be defined on the cross product $M_1 \times M_2 = \{(\bar{x}, \bar{y}) : \bar{x} \in M_1, \bar{y} \in M_2\}$, as:

$$L(\Theta_{2HT}|X,Y) = \begin{cases} \prod_{i}^{M_1} f_{\Theta_{2HT}}(x_i) = (D + (1-D)\gamma)^{\bar{x}}((1-D)\gamma)^{M_1 - \bar{x}} & (3) \\ \prod_{j}^{M_2} f_{\Theta_{2HT}}(y_j) = (D + (1-D)(1-\gamma))^{\bar{y}}((1-D)\gamma)^{M_2 - \bar{y}} & (4) \end{cases}$$

Note, the two Bernoulli random variables should not be confused as bivariate Bernoulli random variables that take values from $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$ in the Cartesian product

space $\{0,1\}^2$. Doing so would confuse the reader into mistakenly assuming the random vectors are inseparable and thus analysis on the likelihood function may be done improperly on the joint probability distribution rather than on each separately. We have expressed the likelihood functions above as piece-wise functions to more clearly elucidate this distinction.

1.5 Jacobian and Hessian

The definitions above are obtained straight from the theoretical assumptions made by each model on a particular set of data for an individual. Given the likelihood functions we are now able to obtain the corresponding maximum likelihood estimators and their range over the cross-product. Since it is easier to work with the log likelihood we move forward with taking the nth-order partial derivative of the log likelihood function. In order to condense our notation, when referring to the likelihood function, $L_{X,Y}(\Theta)$, we mean the $log(L(\Theta|X,Y))$.

Proposition 1.a There exists a unique $\Theta_{SDT} = \{d', \tau\} \in \Omega_{SDT}$ where $\Omega_{SDT} = \{(d', \tau) : d' \in [0,\infty), \tau \in (-\infty,\infty)\}$ $\forall$ $\bar{x}$, $\bar{y}$ if and only if $\bar{x} > 0, \bar{y} > 0$, in $C = \{(\bar{x},\bar{y}) : \bar{x} - \bar{y} > 0\}$

Proof. Using definition 1.a and letting $L_{X,Y}(\Theta_{SDT})$ be the likelihood function differentiable on Θ_{SDT}:

$$\frac{\partial}{\partial d'} L_{X,Y}(\Theta) = \begin{cases} \frac{\partial}{\partial d'} log(f_{\Theta_{SDT}}(\bar{x})) \\ \frac{\partial}{\partial d'} log(f_{\Theta_{SDT}}(\bar{y})) \end{cases}$$

$$\frac{\partial}{\partial\tau}L_{X,Y}(\Theta) = \begin{cases} \frac{\partial}{\partial\tau}log(f_{\Theta_{SDT}}(\bar{x})) \\ \frac{\partial}{\partial\tau}log(f_{\Theta_{SDT}}(\bar{y})) \end{cases}$$

then deriving the score functions and setting them to zero we obtain :

$$= \begin{cases} [M_1\Phi(\frac{d'}{2} - \tau) - \bar{x}] = 0 \\ [M_2\Phi(\frac{-d'}{2} - \tau) - \bar{y}] = 0 \end{cases}$$

Now, solving the above for d' and τ gives:

$$\hat{d'} = \Phi^{-1}(\frac{\bar{x}}{M_1}) - \Phi^{-1}(\frac{\bar{y}}{M_2}) \tag{1.1}$$

$$\hat{\tau} = -\frac{1}{2}[\Phi^{-1}(\frac{\bar{x}}{M_1}) + \Phi^{-1}(\frac{\bar{y}}{M_2})] \tag{1.2}$$

□

where Φ^{-1} is the inverse of the normal distribution function and is equivalent to a z transformation.

Proposition 1.b There exists a unique $\Theta_{2HT} = \{D, \gamma\} \in \Omega_{2HT}$ where $\Omega_{2HT} = \{(D, \gamma) : D \in [0, 1], \gamma \in [0, 1]\}$ $\forall$ $\bar{x}$, $\bar{y}$ if and only if $\bar{x} > 0, \bar{y} > 0$, in $C = \{(\bar{x}, \bar{y}) : \bar{x} - \bar{y} > 0\}$

Proof. Using definition 1.b and letting $L_{X,Y}(\Theta_{2HT})$ be the likelihood function differentiable on Θ_{2HT}:

$$\frac{\partial}{\partial D}L_{X,Y}(\Theta) = \begin{cases} \frac{\partial}{\partial D}log(f_{\Theta_{2HT}}(\bar{x})) \\ \frac{\partial}{\partial D}log(f_{\Theta_{2HT}}(\bar{y})) \end{cases}$$

$$\frac{\partial}{\partial \gamma}L_{X,Y}(\Theta) = \begin{cases} \frac{\partial}{\partial \gamma}log(f_{\Theta_{2HT}}(\bar{x})) \\ \frac{\partial}{\partial \gamma}log(f_{\Theta_{2HT}}(\bar{y})) \end{cases}$$

then the score functions are (respectively):

$$= \begin{cases} (D(-1+\gamma)M_1 - \gamma M_1 + \bar{x}) = 0 \\ ((-1+D)\gamma M_2 + \bar{y}) = 0 \end{cases}$$

$$= \begin{cases} (\bar{x} - DM_1 - \gamma M_1 + D\gamma M_1) = 0 \\ (\bar{y} - \gamma M_2 + D\gamma M_2) = 0 \end{cases}$$

Solving (9) and (10) for D and γ gives:

$$\hat{D} = \frac{\bar{x}}{M_1} - \frac{\bar{y}}{M_2} \tag{1.3}$$

$$\hat{g} = \frac{\frac{\bar{y}}{M_2}}{1 - (\frac{\bar{x}}{M_1}) + (\frac{\bar{y}}{M_2})} \tag{1.4}$$

□

The solutions above are unique and thus we have presented that there exists an extremum in the interior which is specified by setting the first derivative to zero. However, it may be the case that the inflection points found are for a minimum and furthermore it is even possible the solution is only a local extremum and not a global extremum. To determine whether we have found a global maximum, we rely on the second-order partial derivative test using the eigenvalues of the function's Hessian matrix at the critical point. While this test provides evidence for the direction of the curvature of the function at the critical point (minimum v.s. maximum) it does not determine whether the inflection point represents a global or local extremum; a test to check if they are also global points will follow.

Observation: If $L_{X,Y}(\Theta_{SDT})$ is twice differentiable at the critical points $\Theta_{SDT} = \{d', \tau\}$ over $\mathbf{C}$ then the Hessian Matrices $\mathbf{H}$ of $L_{X,Y}(\Theta_{SDT})$ are $\mathbf{H}(L_X(\Theta_{SDT}))$ and $\mathbf{H}(L_Y(\Theta_{SDT}))$ [3]:

$$\mathbf{H}(L_X(\Theta_{SDT})) = \begin{bmatrix} \frac{\partial^2}{\partial d'^2} L_X(\Theta_{SDT}) & \frac{\partial^2}{\partial d' \partial \tau} L_X(\Theta_{SDT}) \\ \frac{\partial^2}{\partial \tau \partial d'} L_X(\Theta_{SDT}) & \frac{\partial^2}{\partial \tau^2} L_X(\Theta_{SDT}) \end{bmatrix} \tag{1.5}$$

$$\mathbf{H}(L_Y(\Theta_{SDT})) = \begin{bmatrix} \frac{\partial^2}{\partial d'^2} L_Y(\Theta_{SDT}) & \frac{\partial^2}{\partial d' \partial \tau} L_Y(\Theta_{SDT}) \\ \frac{\partial^2}{\partial \tau \partial d'} L_Y(\Theta_{SDT}) & \frac{\partial^2}{\partial \tau^2} L_Y(\Theta_{SDT}) \end{bmatrix} \tag{1.6}$$

[3]Once again all operations are done on the individual functions since X, Y are independent.

where the resulting matrix is symmetric, $\frac{\partial^2}{\partial d' \partial \tau} L(\Theta_{SDT}) = \frac{\partial^2}{\partial \tau \partial d'} L(\Theta_{SDT})$ and since from Clairaut's theorem the order of diffferentiation does not matter if the second-order partial derivatives of L are continuous we solve for one of the diagonal second-order partial derivatives.

For the Hessian matrix of the function of X, let $A_1 = e^{(-d'^2/8)}$, $B_1 = e^{(-\tau^2/2)}$, $C_1 = e^{((d'\tau)/2)}$, $D_1 = e^{(-(\frac{(d'-2\tau)}{(2\sqrt{2})})^2)}$, $\Phi_X = \Phi(\frac{d'}{2} - \tau)$

$$\frac{\partial^2}{\partial d'^2} L(\Theta_{SDT}|X) =$$
$$- \frac{A_1 B_1 C_1 (\sqrt{2\pi}(\Phi(\Phi_X - 1)\Phi_X(d' - 2\tau)(\Phi_X M_1 - 1) + 2(2\Phi_X M_1 - 2)))}{16\pi(\Phi_X - 1)^2 \Phi_X}$$

$$\frac{\partial^2}{\partial d' \partial \tau} L_X(\Theta_{SDT}) = \frac{\partial^2}{\partial \tau \partial d'} L_X(\Theta_{SDT}) =$$
$$\frac{A_1 B_1 C_1 (\sqrt{2\pi}(\Phi_X - 1)\Phi_X(d' - 2\tau)(\Phi_X M_1 - 1) + 2(2\Phi_X M_1 - 2))}{8\pi(\Phi_X - 1)^2 \Phi_X}$$

$$\frac{\partial^2}{\partial \tau^2} L_X(\Theta_{SDT}) =$$
$$- \frac{A_1 B_1 C_1 (\sqrt{2\pi}(\Phi_X - 1)\Phi_X(d' - 2\tau)(\Phi_X M_1 - 1) + 2(2\Phi_X M_1 - 2))}{4\pi(\Phi_X - 1)^2 \Phi_X}$$

For the Hessian matrix of the function of Y, let $A_2 = e^{(-2(d'/4+\tau/2)^2)}$, $B_2 = e^{(-(d'/2+\tau)^{2/2})}$ and $\Phi_Y = \Phi(\frac{d'}{2} + \tau)$

$$\frac{\partial^2}{\partial d'^2} L_Y(\Theta_{SDT}) =$$
$$- \frac{B_2(2A_2(\Phi_Y - 1)^2 M_2 + (2\Phi_Y - 1)\bar{y} + \sqrt{2\pi}(\Phi_Y - 1)\Phi_Y(d + 2\tau)((\Phi_Y - 1)M_2 + \bar{y}))}{16\pi(\Phi_Y - 1)^2 \Phi_Y^2}$$

$$\frac{\partial^2}{\partial d' \partial \tau} L_Y(\Theta_{SDT}) = \frac{\partial^2}{\partial \tau \partial d'} L_Y(\Theta_{SDT}) =$$
$$- \frac{B_2(2A_2(\Phi_Y - 1)^2 M_2 + (2\Phi_Y - 1 - 1)y + \sqrt{2\pi}(\Phi_Y - 1 - 1)\Phi_Y - 1(d' + 2\tau)((\Phi_Y - 1)M_2 + y)}{8\pi(\Phi_Y - 1 - 1)^2 \Phi_Y - 1^2}$$

$$\frac{\partial^2}{\partial \tau^2} L_Y(\Theta_{SDT}) =$$
$$- \frac{B_2(2A_2(\Phi_Y - 1)^2 M_2 + (2\Phi_Y - 1 - 1)y + \sqrt{2\pi}(\Phi_Y - 1 - 1)\Phi_Y - 1(d' + 2\tau)((\Phi_Y - 1 - 1)M_2 + y)}{4\pi(\Phi_Y - 1 - 1)^2 \Phi_Y - 1^2}$$

Now, since the determinant of a matrix is the product of the eigenvalues of that matrix, we calculate the determinant of the Hessian matrix first for each Hessian matrix. If the results are positive, it means that the matrix is either positive definite or negative definite and thus either a minimum or maximum, respectively, has been found. If, however, the results are negative or zero, the test will show that we have reached a saddle point or that it is inconclusive, respectively. In our case, the Hessian matrix is singular so $det(\mathbf{H}(L_X(\Theta_{SDT}))) = det(\mathbf{H}(L_Y(\Theta_{SDT}))) = 0$ so this test is inconclusive, meaning that we may have found a minimum, maximum, or saddle point. We will explore this further when checking to see if the MLEs are global extremums.

Observation: If $L_{X,Y}(\Theta_{2HT})$ is twice differentiable at the critical points $\Theta_{2HT} = \{D, \gamma\}$ over $\mathbf{C}$ then the Hessian Matrix $\mathbf{H}$ can also be analyzed separately for each random variable:

$$\mathbf{H}(L_X(\Theta_{2HT})) = \begin{bmatrix} \frac{\partial^2}{\partial D^2} L_X(\Theta_{2HT}) & \frac{\partial^2}{\partial D \partial \gamma} L_X(\Theta_{2HT}) \\ \frac{\partial^2}{\partial \gamma \partial D} L_X(\Theta_{2HT}) & \frac{\partial^2}{\partial \gamma^2} L_X(\Theta_{2HT}) \end{bmatrix} \tag{1.7}$$

$$\mathbf{H}(L_Y(\Theta_{2HT})) = \begin{bmatrix} \frac{\partial^2}{\partial D^2} L_Y(\Theta_{2HT}) & \frac{\partial^2}{\partial D \partial \gamma} L_Y(\Theta_{2HT}) \\ \frac{\partial^2}{\partial \gamma \partial D} L_Y(\Theta_{2HT}) & \frac{\partial^2}{\partial \gamma^2} L_Y(\Theta_{2HT}) \end{bmatrix} \tag{1.8}$$

again the resulting matrix is symmetric, $\frac{\partial^2}{\partial D \partial \gamma} L(\Theta_{2HT}) = \frac{\partial^2}{\partial \gamma \partial D} L(\Theta_{2HT})$:

$$\frac{\partial^2}{\partial D^2} L_X(\Theta_{2HT}) = \frac{1 - M_1}{(D-1)^2} - \frac{(g-1)^2}{(D + (1-D)\gamma)^2}$$

$$\frac{\partial^2}{\partial D \partial \gamma} L_X(\Theta_{2HT}) = \frac{\partial^2}{\partial \gamma \partial D} L_X(\Theta_{2HT}) = \frac{1}{(D + (1-D)\gamma)^2}$$

$$\frac{\partial^2}{\partial \gamma^2} L_X(\Theta_{2HT}) = \frac{1 - M_1}{(D-1)^2} - \frac{(D-1)^2}{(D + (1-D)\gamma)^2}$$

and the elements in the second Hessian matrix are:

$$\frac{\partial^2}{\partial D^2} L_Y(\Theta_{2HT}) = -\frac{\bar{y}}{(D-1)^2} - \frac{g^2(M_2 - \bar{y})}{(1 + (D-1)\gamma)^2}$$

$$\frac{\partial^2}{\partial D \partial \gamma} L_Y(\Theta_{2HT}) = \frac{\partial^2}{\partial \gamma \partial D} L_Y(\Theta_{2HT}) = \frac{M_2 - \bar{y}}{(1 + (D-1)\gamma)^2}$$

$$\frac{\partial^2}{\partial \gamma^2} L_Y(\Theta_{2HT}) = -\frac{\bar{y}}{\gamma^2} - \frac{(D-1)^2(M_2 - \bar{y})}{(1 + (D-1)\gamma)^2}$$

The results show that both $det(\mathbf{H}(\mathbf{L}(\mathbf{\Theta_{2HT}}|\mathbf{X}))) > 0$; and $det(\mathbf{H}(\mathbf{L}(\mathbf{\Theta_{2HT}}|\mathbf{Y}))) > 0$ thus our results for the 2HT are either maximum or minumum points and not saddle points. Finding the eigenvalues of each matrix reveals that the Hessian matrix is negative definite; thus we have found maximums.

To verify that the MLE is the global maximum we test the end points $\{x, y\}$ are either $\{0, 0\}$ or $\{M_1, M_2\}$.

$$L(SDT) = \begin{cases} M_1 ln(\Phi(-\frac{\hat{d'}}{2} + \hat{\tau})) + M_2 ln(\Phi(\frac{\hat{d'}}{2} + \hat{\tau})) & \text{if } x = 0 \text{ and } y = 0 \\ M_1 ln(\Phi(\frac{\hat{d'}}{2} - \hat{\tau})) + M_2 ln(\Phi(\frac{\hat{d'}}{2} + \tau)) & \text{if } x = M_1 \text{ and } y = 0 \\ M_1 ln(\Phi(-\frac{\hat{d'}}{2} + \hat{\tau})) + M_2 ln(\Phi(-\frac{\hat{d'}}{2} - \hat{\tau})) & \text{if } x = 0 \text{ and } y = M_2 \\ M_1 ln(\Phi(\frac{\hat{d'}}{2} - \hat{\tau})) + M_2 ln(\Phi(-\frac{\hat{d'}}{2} - \hat{\tau})) & \text{if } x = M_1 \text{ and } y = M_2 \end{cases}$$

$$L(2HT) = \begin{cases} M_1 ln((1 - \hat{D})(1 - \hat{\hat{\gamma}})) + M_2 ln(\hat{D} + (1 - \hat{D})(1 - \hat{\gamma})) & \text{if } x = 0 \text{ and } y = 0 \\ M_1 ln(\hat{D} + (1 - \hat{D})\hat{\gamma}) + M_2 ln(\hat{D} + (1 - \hat{D})(1 - \hat{\gamma})) & \text{if } x = M_1 \text{ and } y = 0 \\ M_1 ln((1 - \hat{D})(1 - \hat{\gamma})) + M_2 ln((1 - \hat{D})\hat{\gamma}) & \text{if } x = 0 \text{ and } y = M_2 \\ M_1 ln(\hat{D} + (1 - \hat{D})\gamma) + M_2 ln((1 - \hat{D})\hat{\gamma}) & \text{if } x = M_1 \text{ and } y = M_2 \end{cases}$$

It is easily seen that the first set of functions do not readily converge. The difficulty lies in the function Φ^{-1} (contained in the MLE's of SDT) such that whenever the input is either $\{0, 0\}$ or $\{M_1, M_2\}$, the result is either $-\infty$ or ∞, respectively. Restricting the range to the open interval $(0, 1)$ fixes this problem and allows us to see that the earlier result was a local maximum. As for the second set of functions, it is straightforward to verify that the MLE's of 2HT still produce a maximum in each case. Thus the maximum found for the 2HT model is global. For the remainder we will assume global maximums for both models within the restricted open interval $(0, 1)$ in order to proceed with our comparisons.

1.6 Fisher's Information

Now that we have shown where the MLEs are maximums, and have defined their probability space we turn to exploring more about the parameters themselves starting with calculating Fisher's Information. Fisher's Information is prescribed as a method of quantifying a measure of information from an observation of a random variable X about a parameter θ. Theoretically examining this methodology we find that the information obtained is defined as a measure of the variance of the score function. Formally[4]

$$\begin{aligned} I_{XY}(\Theta) &= I_X(\Theta) + I_Y(\Theta) \\ &= E[\frac{\partial}{\partial\theta} log f(X,\theta)^2|\theta] + E[\frac{\partial}{\partial\theta} log f(Y,\theta)^2|\theta] \end{aligned}$$

Since we have seen that the log likelihood is twice differentiable with respect to the parametersfor the models under consideration (SDT and 2HT), we choose to use the alternative form:

$$\begin{aligned} I_{XY}(\Theta) &= I_X(\Theta) + I_Y(\Theta) \\ &= -E[\frac{\partial^2}{\partial\Theta^2} log f(X,\Theta)|\theta] + -E[\frac{\partial^2}{\partial\Theta^2} log f(Y,\Theta)|\theta] \end{aligned}$$

For the current models we can write the Fisher's Information as the expected value of the corresponding Hessian matrix given the parameters:

[4]Note, since X and Y are independent, the Fisher's Information of the combined likelihood function is the sum of the individual Fisher's Information since information is additive.

$$I_{XY}(\Theta)_{jj'} = I_X(\Theta)_{jj'} + I_Y(\Theta)_{jj'}$$
$$= -E[\mathbf{H}(L_X(\Theta))|\Theta] + -E[\mathbf{H}(L_Y(\Theta))|\Theta]$$

where $j = 1, \ldots, K$ and K is the number of parameters in the model.

We begin with the Fisher's Information matrices for signal detection theory. Let $A_3 = e^{(-d'^2/4)}$, $B_3 = e^{(-\tau^2)}$, $C_3 = e^{(d'\tau)}$, $D_3 = e^{(-d'\tau)}$ then:

$$\mathbf{I_X}(\mathbf{\Theta_{SDT}}) = \begin{bmatrix} -\frac{A_3B_3C_3M_1}{2\pi((2\Phi_X-1)^2-1)} & \frac{A_3B_3C_3M_1}{\pi((2\Phi_X-1)^2-1)} \\ \frac{A_3B_3C_3M_1}{\pi((2\Phi_X-1)^2-1)} & -\frac{2A_3B_3C_3M_1}{\pi((2\Phi_X-1)^2-1)} \end{bmatrix} \tag{1.9}$$

$$\mathbf{I_Y}(\mathbf{\Theta_{SDT}}) = \begin{bmatrix} -\frac{A_3B_3D_3M_2}{2\pi((2\Phi_Y-1)^2-1)} & -\frac{A_3B_3D_3M_2}{\pi((2\Phi_Y-1)^2-1)} \\ -\frac{A_3B_3D_3M_2}{\pi((2\Phi_Y-1)^2-1)} & -\frac{2A_3B_3D_3M_2}{\pi((2\Phi_Y-1)^2-1)} \end{bmatrix} \tag{1.10}$$

The Fisher's Information for the double high threshold model is:

$$\mathbf{I_X}(\mathbf{\Theta_{2HT}}) = \begin{bmatrix} \frac{(1-g)M_1}{(1-D)(D+(1-D)g)} & \frac{M_1}{D+(1-D)g} \\ \frac{M_1}{D+(1-D)g} & \frac{(1-D)M_1}{(-D-(1-D)g)(-1+g)} \end{bmatrix} \tag{1.11}$$

$$\mathbf{I_Y}(\mathbf{\Theta_{2HT}}) = \begin{bmatrix} -\frac{gM_2}{(D-1)(1-(1-D)g)} & -\frac{M_2}{1+(D-1)g} \\ -\frac{M_2}{1+(D-1)g} & \frac{(1-D)M_2}{(g(1+(D-1)g)} \end{bmatrix} \tag{1.12}$$

The Fisher's Information for each of the models is then the sum of the two independent Fisher's Information matrices. The solutions provide us with the amount of information each random variable carries about the parameters of the model.

1.6.1 Parameter Variances

The precision to which we can estimate the parameters can be calculated by an inverse transformation on the Fisher's Information. For example, with one parameter:

$$V(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

However, this transformation for more than one parameter is not always simply the inverse of the individual diagonal elements in the Fisher's Information matrix (e.g. when the parameters in the model are not orthogonal). Instead, we invert the Fisher's Information matrix and take results from the diagonal elements of the inverted matrix for the corresponding parameters of interest. Formally:

$$Cov(\hat{\Theta}) \geq I(\Theta)^{-1}_{jj'} \tag{1.13}$$

The resulting matrix is a matrix with parameter variances on the main diagonal and covariances on the off-diagonal elements.

For the SDT model, the variances and covariances are:

$$Var(\hat{d}') = -\frac{(2\pi((-1+G)GM_2e^{1/4(d-2t)^2} + (-1+E)EM_1e^{1/4(d+2t)^2}))}{M_1M_2} \tag{1.14}$$

$$Var(\hat{\tau}) = -\frac{(\pi((-1+G)GM_2e^{1/4(d-2\tau)^2} + (-1+E)EM_1e^{1/4(d+2\tau)^2}))}{2M_1M_2} \tag{1.15}$$

$$cov(\hat{d}', \hat{\tau}) = cov(\hat{\tau}, \hat{d}') = \frac{\pi((-1+G)GM_2e^{1/4(d-2t)^2} - (-1+E)EM_1e^{1/4(d+2t)^2})}{M_1M_2}$$

For the 2HT model the variances and covariances are:

$$Var(\hat{D}) = -\frac{(-1+D)(-(-1+\gamma)\gamma(M_2+M_1) + D((-1+\gamma)^2M_2 + \gamma^2M_1))}{M_1M_2} \tag{1.16}$$

$$Var(\hat{\gamma}) = \frac{(-1+\gamma)\gamma(M_1 - 2\gamma M_1 - D(-1+\gamma)\gamma(M_1+M_2) + \gamma^2(M_1+M_2))}{((M_1M_2(D-1)))} \tag{1.17}$$

$$Cov(\hat{D}, \hat{\gamma}) = Cov(\hat{\gamma}, \hat{D}) = \frac{(-1+\gamma)\gamma(M_1 - \gamma(M_1+M_2) + D((-1+\gamma)M_2 + \gamma M_1))}{(M_1M_2)}$$

The results reveal that the variances of both parameters of the 2HT model are much smaller than for the parameters of SDT. With the variances of the estimators in hand, testing consistency is possible by examining the limiting properties of the variance as M_1 and M_2 go to infinity. The general form is:

$$\lim_{M_1 \to \infty} \lim_{M_2 \to \infty} Var(\hat{\theta}) \tag{1.18}$$

The results show the estimators as being consistent such that when both M_1 and M_2 go to infinity the variance decreases to zero for both models.

1.7 Parameter sensitivity and predictions

Often the predictions made by the model parameters are put into question through selective influence studies that vary a particular aspect of an experiment to show the sensitivity of the parameters to changes in behavior. Many times the rigorous testing of parameters and their predictions leads to an understanding of psychological phenomena. However, the data gathered by psychologists often contain systematic errors that can have unwanted consequences, such as numerical instability where small changes in the data may have an outsized effect on our resulting conclusions.

To calculate the effect small changes in the data may have on the parameter estimates, we make the assumption that the data is measured with some error, ϵ_x and ϵ_y for each random variable X and Y respectively. The new likelihood function that expresses this belief is $L(\Theta|X + \epsilon_x, Y + \epsilon_y)$. We would expect the new MLEs to be close to the uncorrected MLEs when ϵ_x and ϵ_y are small.

The MLEs for the SDT parameters are:

$$\bar{d'} = \Phi^{-1}(\frac{\bar{x} + \epsilon_x}{M_1}) - \Phi^{-1}(\frac{\bar{y} + \epsilon_y}{M_2}) \tag{1.19}$$

$$\bar{\tau} = -\frac{1}{2}[\Phi^{-1}(\frac{\bar{x} + \epsilon_x}{M_1}) + \Phi^{-1}(\frac{\bar{y} + \epsilon_y}{M_2})] \tag{1.20}$$

The MLEs for the 2HT parameters are:

$$\bar{D} = \frac{\bar{x} + \epsilon_x}{M_1} - \frac{\bar{y} + \epsilon_y}{M_2} \tag{1.21}$$

$$\bar{g} = \frac{\frac{\bar{y}+\epsilon_y}{M_2}}{1 - (\frac{\bar{x}+\epsilon_x}{M_1}) + (\frac{\bar{y}+\epsilon_y}{M_2})} \tag{1.22}$$

The new MLEs are very similar to those above, except they now include the error variable. If we posit that errors in the data are small, i.e. $\epsilon_x = 1$ and $\epsilon_y = 1$, we can plot the expected value in such a way that would show us how the parameters vary for all $\bar{x}$ and $\bar{y}$. In order to simplify the graphical display we limit our exploration to the recursive relation:

$$\bar{x}_t = \bar{x}_{t-1} + \epsilon_x \tag{1.23}$$

$$\bar{y}_t = \bar{y}_{t-1} + \epsilon_y \tag{1.24}$$

such that when $\bar{x}_0 = 1$ and $\bar{y}_0 = 0$ the difference is 1. By adding a constant error value of 1 to each iteration, it is obvious that the resulting $\bar{x}_t$ and $\bar{y}_t$ will retain a difference of 1.

Figure 3 is obtained by replacing (1.23) and (1.24) into equations 1.19-1.22 with an arbitrary sample size of $M_1 = M_2 = 50$.

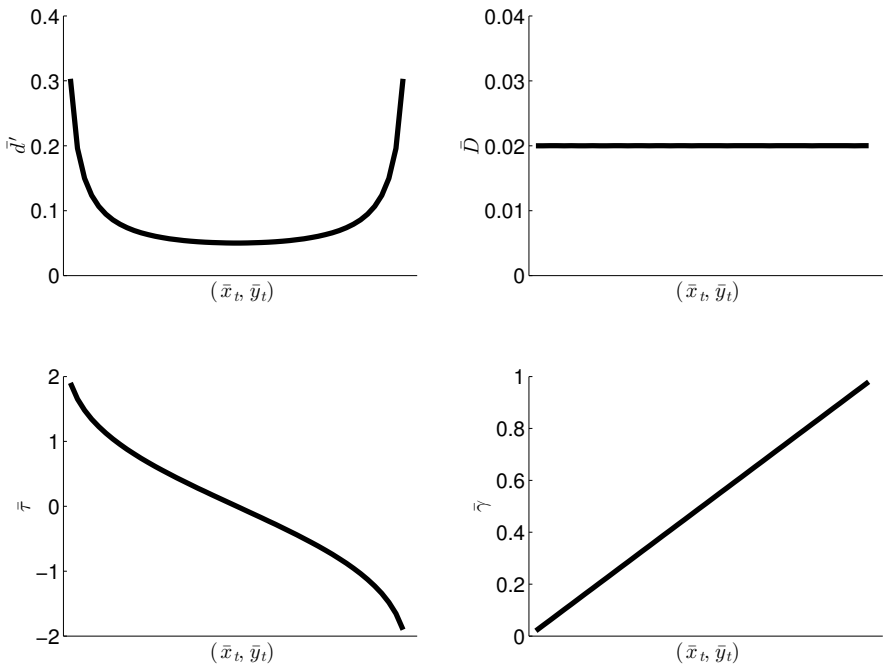

Figure 1.3: Parameter Sensitivity Test (Cross Section)

Figure 3 shows the cross-section of how sensitive the parameters are across the possible success quantities when the error is held fixed at 1. For example, the figure shows how sensitive d' is to small and large values of $\bar{x}$ and $\bar{y}$ compared to middle values. Unlike d', D shows no change with small changes in the data.

Now, a point of contention by proponents of SDT is the mistrust of threshold models being used as measurement models regardless of their validity as psychological models (Pazzaglia, Dube, and Rotello, 2013). In an effort to discredit threshold models, Pazzaglia et al. argued that disagreement between the models may arise and can often lead researchers to misinterpret data. However, the disagreement in question comes from misidentifying the latent variables associated with one model, such as memory ability, as the leading difference between data from two groups of subjects [Harvey; Kinshla]. When analyzing each group with

the two models, it is noted that a change in memory sensitivity is predicted by one model while the other reports it as a change in bias.

Thus far we have looked at the models when a single data set from one person has been collected. Although this has generally given researchers sufficient information to apply the models for the study of memory, it limits the types of model comparison tests. The discrepancies reported by Pazzaglia et al., if true, constitute an important finding for the goal of determining the best model for psychological theories of memory. While that is not our goal here, we explore the comparison between predictions made by each model for two subjects with different response patterns. This is akin to the well-known between subject studies.

Let $X^{(1)}$, $X^{(2)}$, $Y^{(1)}$ and $Y^{(2)}$ be discrete independent bivariate random vectors from two subjects with success probabilities $\{p^{(1)}, p^{(2)}, q^{(1)}, q^{(2)}\}$, corresponding to four sequences of independent Bernoulli trials with lengths M. Denote the quadruple product space for the number of success trials as $\Omega_E = \{(\bar{x}_I, \bar{x}_{II}, \bar{y}_I, \bar{y}_{II}) : \bar{x}_I \in A^{(1)}, \bar{x}_{II} \in A^{(2)}, \bar{y}_I \in B^{(1)}, \bar{y}_{II} \in B^{(2)}\}$ where $A^{(1)} \subset \mathbb{N}$, $A^{(2)} \subset \mathbb{N}$, $B^{(1)} \subset \mathbb{N}$, and $B^{(2)} \subset \mathbb{N}$. Then the complete sample space Ω_E can be partitioned into four sets of equal cardinality by:

$$\Omega_{E_1} = \{(\bar{x}_I, \bar{x}_{II}, \bar{y}_I, \bar{y}_{II}) : \bar{x}_I \geq \bar{x}_{II}, \bar{y}_I \leq \bar{y}_{II}\}$$
$$\Omega_{E_2} = \{(\bar{x}_I, \bar{x}_{II}, \bar{y}_I, \bar{y}_{II}) : \bar{x}_I \leq \bar{x}_{II}, \bar{y}_I \geq \bar{y}_{II}\}$$
$$\Omega_{E_3} = \{(\bar{x}_I, \bar{x}_{II}, \bar{y}_I, \bar{y}_{II}) : \bar{x}_I \geq \bar{x}_{II}, \bar{y}_I \geq \bar{y}_{II}\}$$
$$\Omega_{E_4} = \{(\bar{x}_I, \bar{x}_{II}, \bar{y}_I, \bar{y}_{II}) : \bar{x}_I \leq \bar{x}_{II}, \bar{y}_I \leq \bar{y}_{II}\}$$

The first two sets, Ω_{E_1} and Ω_{E_2}, characterized by linear inequalities on the sample space Ω_E, describe the well-known mirror effect and account for a majority of the psychological data gathered for two group experiments (Iverson, Glanzer et al.). The other two sets are not frequently observed in psychological studies, and we will show that disagreement between the models lies within these regions.

Theorem 1.1. *For any quadruple from the sets Ω_{E_1} and Ω_{E_2} the following differences $D^{(1,2)} = D^{(1)} - D^{(2)}$ and $d'^{(1,2)} = d'^{(1)} - d'^{(2)}$ have the same parity. Furthermore, for any quadruple from the sets Ω_{E_3} and Ω_{E_4} the following differences $D^{(1,2)} = D^{(1)} - D^{(2)}$ and $d'^{(1,2)} = d'^{(1)} - d'^{(2)}$ contain instances with opposite parity.*

Proof. First, assume a probability quartet, $\{p^{(1)}, p^{(2)}, q^{(1)}, q^{(2)}\}$, satisfying the linear inequalities from Ω_{E_1}. Then from Definition 1 and without loss of generality:

$$\begin{aligned} p^{(1)} > p^{(2)} &= \Phi(\frac{d'^{(1)}}{2} - \tau^{(1)}) > \Phi(\frac{d'^{(2)}}{2} - \tau^{(2)}) \\ &= \frac{d'^{(1)}}{2} - \tau^{(1)} > \frac{d'^{(2)}}{2} - \tau^{(2)} \end{aligned}$$

$$\frac{d'^{(1)}}{2} - \frac{d'^{(2)}}{2} > \tau^{(1)} - \tau^{(2)} \tag{1.25}$$

$$\begin{aligned} q^{(1)} < q^{(2)} &= \Phi(-\frac{d'^{(1)}}{2} - \tau^{(1)}) < \Phi(-\frac{d'^{(2)}}{2} - \tau^{(2)}) \\ &= \frac{-d'^{(1)}}{2} - \tau^{(1)} < \frac{-d'^{(2)}}{2} - \tau^{(2)} \end{aligned}$$

$$\frac{d'^{(1)}}{2} - \frac{d'^{(2)}}{2} > \tau^{(2)} - \tau^{(1)} \tag{1.26}$$

and by Definition 2:

$$p^{(1)} > p^{(2)} = D^{(1)} + (1 - D^{(1)})\gamma^{(1)} > D^{(2)} + (1 - D^{(2)})\gamma^{(2)}$$

$$D^{(1)} - D^{(2)} > (1 - D^{(2)})\gamma^{(2)} - (1 - D^{(1)})\gamma^{(1)} \tag{1.27}$$

$$q^{(1)} < q^{(2)} = (1 - D^{(1)})\gamma^{(1)} < (1 - D^{(2)})\gamma^{(2)}$$

$$(1 - D^{(2)})\gamma^{(2)} - (1 - D^{(1)})\gamma^{(1)} > 0 \tag{1.28}$$

Next, assume a probability quartet, $\{p^{(1)}, p^{(2)}, q^{(1)}, q^{(2)}\}$, satisfying the linear inequalities from Ω_{E_3}. Once again from Definition 1 and without loss of generality:

$$\begin{aligned} p^{(1)} > p^{(2)} &= \Phi(\frac{d'^{(1)}}{2} - \tau^{(1)}) > \Phi(\frac{d'^{(2)}}{2} - \tau^{(2)}) \\ &= \frac{d'^{(1)}}{2} - \tau^{(1)} > \frac{d'^{(2)}}{2} - \tau^{(2)} \end{aligned}$$

$$\frac{d'^{(1)}}{2} - \frac{d'^{(2)}}{2} > \tau^{(1)} - \tau^{(2)} \tag{1.29}$$

$$\begin{aligned} q^{(1)} > q^{(2)} &= \Phi(-\frac{d'^{(1)}}{2} - \tau^{(1)}) > \Phi(-\frac{d'^{(2)}}{2} - \tau^{(2)}) \\ &= \frac{-d'^{(1)}}{2} - \tau^{(1)} > \frac{-d'^{(2)}}{2} - \tau^{(2)} \end{aligned}$$

$$\frac{d'^{(2)}}{2} - \frac{d'^{(1)}}{2} > \tau^{(1)} - \tau^{(2)} \tag{1.30}$$

(2HT)

$$p^{(1)} > p^{(2)} = D^{(1)} + (1 - D^{(1)})\gamma^{(1)} > D^{(2)} + (1 - D^{(2)})\gamma^{(2)}$$

$$D^{(1)} - D^{(2)} > (1 - D^{(2)})\gamma^{(2)} - (1 - D^{(1)})\gamma^{(1)} \tag{1.31}$$

$$q^{(1)} > q^{(2)} = (1 - D^{(1)})\gamma^{(1)} > (1 - D^{(2)})\gamma^{(2)}$$

$$0 > (1 - D^{(2)})\gamma^{(2)} - (1 - D^{(1)})\gamma^{(1)} \tag{1.32}$$

□

It follows from solutions 1.25 and 1.26, $d'^{(1)} > d'^{(2)}$ and by solutions 1.27 and 1.28, $D^{(1)} > D^{(2)}$, therefore $D^{(1,2)}$ and $d'^{(1,2)}$ have the same parity. Furthermore, it is easy to see that when the inequalities are reversed, all the while retaining the mirror effect as in the subset Ω_{E_2}, the results hold. Now, by equations 1.29 and 1.30, $d'^{(1)}$ may be larger or smaller than $d'^{(1)}$. Similarly, by equations 1.31 and 1.32, $D^{(1)}$ may be larger or smaller than $D^{(2)}$. Under the current results, it is not justifiable to assume falsification of one model may be found in a two group study.

To further ascertain when the two models are in discord, we obtain all hit rates and false

alarm rates for two groups, $\{P_{o1}^{(1)},\ P_{n1}^{(1)}\}$ and $\{P_{o1}^{(2)},P_{n1}^{(2)}\}$, using the same sample sizes as before. The simulated probabilities have the same property as those obtained using a two group experimental design, if all data patterns were to be observed. We first calculate a pair of parameters $\{D_i, D_j\}$ and $\{d_i', d_j'\}$ where $i \neq j$, from a common set of observation probabilities and compare their differences for consistency using a sign test. For every probability distribution in each group, the parameters D and d' are calculated from their closed form expressions. Agreement is met when $D_i - D_j > 0$ and $d_i' - d_j' > 0$ or $D_i - D_j < 0$ and $d_i' - d_j' < 0$. Furthermore, we explore the cases when the data values are restricted within $(1-a, a)$ where $a = 1, .9, and .8$.

Table 1.1: Agreement

N	$a = 1$	$a = .9$	$a = .8$
25	92.83%	95.67 %	95.71%
50	92.46%	96.53%	97.78%
100	92.14%	96.69%	98.29%
200	91.94%	96.73%	98.44%

Results of the sign test are presented in Table 1 and it shows agreement between the parameters at over 90% for the four sample sizes. Moreover, when P_o and P_n rates are restricted to non-extreme values, the agreement increases to over 98%. Although the disagreement is small between the models, it cannot be ignored. It must be noted that the psychological predictions are not necessarily predicting different psychological outcomes but rather, because the parameters are not 1-to-1 functions of each other, their psychological interpretation cannot be either. For example, whenever d' is greater in one group, it means that their memory strength is greater compared to the noise distribution, while a smaller D for that group does not mean a diminished memory strength, but rather a diminished ability to detect that memory signal.

1.7.1 Coinciding Likelihoods

We conclude our comparison of the two models by showing that although the two have very different analytical forms and theoretical descriptions their likelihood functions coincide. First it is easy to see that without the inequality constraint, $\bar{x} \geq \bar{y}$, neither model would have valid MLE estimates within their respective theoretical distributions. While these constraints are a product of theoretical assumptions rather than mathematical restrictions, propositions 1.a and 1.b demonstrate that both models entail the same exact set of probability distributions, C, on the sample space of the data. In other words, the two models are statistically equivalent or statistically indistinguishable in a product binomial data structure (Batchelder & Alexander, 2013).

To further illustrate this fact, we evoke the invariance property of MLEs such that: If $\hat{\Theta}$ is the MLE of Θ, then for any function $T(\Theta) = \eta$ the MLE of $T(\Theta)$ is $T(\hat{\Theta})$ and thus the induced likelihood function, $L^*(\eta|x)$ is defined by:

$$L^*(\eta|x) = \sup_{\{\Theta:T(\Theta)=\eta\}} L(\Theta|x)$$

Lemma 1.1. *For any $x > 0, y > 0$, in $C = \{(x, y) : x - y > 0\}$, and $\Theta_{SDT} = (d', \tau)$ where $\Theta_{SDT} \in \Omega_{SDT}$, there exists a pair of functions $T = \{h, g\}$ such that $h(\Theta_{SDT}) = \Phi(\frac{d'}{2} - \tau) - \Phi(-\frac{d'}{2} - \tau) = \eta_D$ and $g(\Theta_{SDT}) = \frac{\Phi(-\frac{d'}{2}-\tau)}{1-\Phi(\frac{d'}{2}-\tau)+\Phi(-\frac{d'}{2}-\tau)} = \eta_\gamma$. Then $\eta_{SDT} = \{\eta_D, \eta_\gamma\}$ where the maxima of the induced likelihood function, $L^*(\eta_{SDT}|x, y)$ and profile likelihood $L(\Theta_{SDT}|x, y)$ coincide. Thus:*

$$L^*(\hat{\eta}_{SDT}|x, y) = L^*(T(\hat{\Theta}_{SDT})|x, y)$$

Proof. Let $\hat{\eta}_{SDT}$ denote the values that maximize $L^*(\eta_{SDT}|x, y)$ and by the definition of the

induced likelihood;

$$\begin{aligned} L^*(\hat{\eta}_{SDT}|x,y) &= \sup_{\eta_{SDT}} \sup_{\{\Theta_{SDT}:T(\Theta_{SDT})=\eta_{SDT}\}} L(\Theta_{SDT}|x,y) \\ &= \sup_{\Theta_{SDT}} L(\Theta_{SDT}|x,y) \\ &= L(\hat{\Theta}_{SDT}|x,y) \end{aligned}$$

Furthermore,

$$\begin{aligned} L(\hat{\Theta}_{SDT}|x,y) &= \sup_{\{\Theta_{SDT}:T(\Theta_{SDT})=T(\hat{\Theta}_{SDT})\}} L(\Theta_{SDT}|x,y) \\ &= L^*(T(\hat{\Theta}_{SDT})|x,y) \end{aligned}$$

Therefore, $L^*(\hat{\eta}_{SDT}|x,y) = L^*(T(\hat{\Theta}_{SDT})|x,y)$

□

Lemma 1.2. *For any $x > 0, y > 0$, in $C = \{(x,y) : x - y > 0\}$, and $\Theta_{2HT} = (D, \gamma)$ where $\Theta_{2HT} \in \Omega_{2HT}$, there exists a pair of functions $T = \{h, g\}$ such that $h(\Theta_{2HT}) = \Phi^{-1}(D + (1-D)\gamma) - \Phi^{-1}((1-D)\gamma) = \eta_{d'}$ and $g(\Theta_{2HT}) = \frac{1}{2}\Phi^{-1}(D + (1-D)\gamma) + \Phi^{-1}((1-D)\gamma) = \eta_{\tau}$ and $\eta_{2HT} = \{\eta_{d'}, \eta_{\tau}\}$ where the maxima of the induced likelihood function, $L^*(\eta_{2HT}|x,y)$ and profile likelihood $L(\Theta_{2HT}|x,y)$ coincide. Thus:*

$$L^*(\hat{\eta}_{2HT}|x,y) = L^*(T(\hat{\Theta}_{2HT})|x,y)$$

Proof. Let $\hat{\eta}_{2HT}$ denote the values that maximizes $L^*(\eta_{2HT}|x,y)$ and by the definition of the

induced likelihood.

$$\begin{aligned} L^*(\hat{\eta}_{2HT}|x,y) &= \sup_{\eta_{2HT}} \sup_{\{\Theta_{2HT}:T(\Theta_{2HT})=\eta_{2HT}\}} L(\Theta_{SDT}|x,y) \\ &= \sup_{\Theta_{2HT}} L(\Theta_{2HT}|x,y) \\ &= L(\hat{\Theta}_{2HT}|x,y) \end{aligned}$$

Furthermore,

$$\begin{aligned} L(\hat{\Theta}_{2HT}|x,y) &= \sup_{\{\Theta_{2HT}:T(\Theta_{2HT})=T(\hat{\Theta}_{2HT})\}} L(\Theta_{2HT}|x,y) \\ &= L^*(T(\hat{\Theta}_{2HT})|x,y) \end{aligned}$$

Therefore, $L^*(\hat{\eta}_{2HT}|x,y) = L^*(T(\hat{\Theta}_{2HT})|x,y)$

□

Theorem 1.2. *For any $x > 0, y > 0$, in $C = \{(x,y) : x - y > 0\}$, the maxima of $L(\Theta_{SDT}|x,y)$ and $L(\Theta_{2HT}|x,y)$ coincide.*

Proof. It is easy to see from Lemma 1 that the maxima of $L(\hat{\Theta}_{SDT}|x,y)$ and $L^*(T(\hat{\Theta}_{SDT})|x,y)$ coincide. To show that the maxima of $L(\Theta_{SDT}|x,y)$ and $L(\Theta_{2HT}|x,y)$ are the same for any $(x,y) \in C$ we first consider the values that maximize $L^*(T(\hat{\Theta}_{SDT})|x,y)$.

From Lemma 1:

$$\hat{\eta}_D = \Phi(\frac{\hat{d'}}{2} - \hat{\tau}) - \Phi(-\frac{\hat{d'}}{2} - \hat{\tau}) \tag{1.33}$$

$$\hat{\eta}_\gamma = \frac{\Phi(-\frac{\hat{d'}}{2} - \hat{\tau})}{1 - \Phi(\frac{\hat{d'}}{2} - \hat{\tau}) + \Phi(-\frac{\hat{d'}}{2} - \tau)} \tag{1.34}$$

Substituting $\{\hat{d'}, \hat{\tau}\}$ with equations 7 and 8 yields:

$$\hat{\eta}_D = H - FA \tag{1.35}$$

$$\hat{\eta}_\gamma = \frac{FA}{1 - HR + FA}. \tag{1.36}$$

Thus we find that the values that maximize $L^*(T(\hat{\Theta}_{SDT})|x, y)$ are equivalent to those in $L(\hat{\Theta}_{2HT}|x, y)$ so $L^*(\eta_{SDT}|x, y) = L(\Theta_{2HT}|x, y)$.

Now, from Lemma 2 we see that the maxima of $L(\hat{\Theta}_{2HT}|x, y)$ and $L^*(T(\hat{\Theta}_{2HT})|x, y)$ coincide as well. Therefore, consider the values that maximize $L^*(T(\hat{\Theta}_{2HT})|x, y)$

$$\hat{\eta}_{d'} = \Phi^{-1}(\hat{D} + (1 - \hat{D})\hat{\gamma}) - \Phi^{-1}((1 - \hat{D})\hat{\gamma}) \tag{1.37}$$

$$\hat{\eta}_\tau = \frac{1}{2}\Phi^{-1}(\hat{D} + (1 - \hat{D})\hat{\gamma}) + \Phi^{-1}((1 - \hat{D})\hat{\gamma}) \tag{1.38}$$

Substituting $\{\hat{D}, \hat{\gamma}\}$ with equations 11 and 12 yields:

$$\hat{\eta}_D = \Phi^{-1}(H) - \Phi^{-1}(FA) \tag{1.39}$$

$$\hat{\eta}_\gamma = -\frac{1}{2}[\Phi^{-1}(H) + \Phi^{-1}(FA)] \tag{1.40}$$

We find that the values that maximize $L^*(T(\hat{\Theta}_{2HT})|x,y)$ are equivalent to those in $L(\hat{\Theta}_{SDT}|x,y)$ thus $L^*(\eta_{2HT}|x,y) = L(\Theta_{SDT}|x,y)$.

Therefore, for any $x > 0, y > 0$, in $C = \{(x,y) : x - y > 0\}$ maximizing either $L(\Theta_{SDT}|x,y)$ or $L(\Theta_{2HT}|x,y)$ will yield the same answer.

□

We can see from the relationship above that the two models are statistically indistinguishable and any comparison between the basic models leading to positive support using methods such as AIC, NML and goodness of fit measures that utilize the fit of the model through the likelihood function will differ only as a result of a penalty term[5] for added complexity. Note, the model complexity term is a function of the parameters and their distributions; it is thus important to be cautious when implementing parametric constraints since the decision to impose theoretical constraints on the parameters as is the case for isosensitivity manipulations will increase the chance of worse fit. Furthermore, it is wise to be careful in the choice of prior distributions for the parameters because it is possible to bias the comparison by assigning improper priors to one model.

1.8 Conclusions

In this paper we sought to formalize the SDT and 2HT models using mathematical convention and succeeded in showing the inner workings of each model. An important property of each model is their likelihood function, for which we are able to compute the maximum likelihood estimators, utilize Bayesian inference techniques, and measure the fit. In our analysis we have shown that the two likelihood functions coincide on every point in the probability

[5]Note, the penalty term does not improve the overall fit of the model but rather it reduces the fit of a model. This means that for any model, the goodness of fit measured by the maximized likelihood function is the best fit attainable by the model.

distribution space and thus show the two models fit the data equally well.

Since the two models are statistically equivalent and methods constructed to compare the two models often involve statistical inference techniques, we did not expect to find differences between the two models at the basic level of goodness of fit. However, mathematically the two models are distinct, an obvious fact from looking at each model's functional form, and from this we hypothesized that differences between the models must exist. From the Hessian matrix, we were able to detail the limitations concerning global maximums and calculate the Fisher's Information measure. The results of this measure instantiated our beliefs that the amount of information each parameter describes for the random variables are outstandingly different. While these results are interesting at a theoretical level, we made an effort to demonstrate the pragmatic value of these results for experimenters.

Finally, parameter sensitivity is an important mathematical analysis for testing numerical instability. Issues of numerical instability are detrimental for research relying on sensitive measures needed to adjudicate the precise treatment. Our results show that SDT parameters are extremely sensitive to small perturbations in the data and thus may lead to misrepresenting the severity of memory deficits.

1.9 References

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. Journal of the Optical Society of America, 53, 129-160.

Bozdogan, H. (2000). Akaikes information criterion and recent developments in information complexity. Journal of Mathematical Psychology, 44, 62-91.

Broder, A., & Schutz, J. (2009). Recognition ROCs are curvilinear - or are they? on premature arguments against the two-high-threshold model of recognition. Journal of Experimental Psychology: Learning, Memory and Cognition, 35, 587-606.

DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 18-33.

Dube, C., Rotello, C. M., & Heit, E. (2011). The belief bias effect is aptly named: a reply to Klauer and Kellen (2011). Psychological Review, 118, 155-163.

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. Journal of Experimental Psychology: Learning, Memory and Cognition, 38, 130-151.

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

Jeffreys, H. (1961). The theory of probability. Oxford: Oxford University Press.

Karabatsos, G., & Walker, S. G. (2006). On the normalized maximum likelihood and Bayesian decision theory. Journal of Mathematical Psychology, 50, 517-520.

Kass, R., & Rafterty, A. (1995). Bayes factors. Journal of the American Statistical Association, 90, 773-795.

Klauer, K. C., & Kellen, D. (2011a). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (2010). Psychological Review, 118, 164-173.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-464.

Swets J A, Tanner W P Jr & Birdsall T G. Decision processes in perception. Psychol. Rev. 68: 301-40, 1961.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. Psychological Review, 114, 152-176.

Chapter 2

Statistical Development and Comparison of two Recognition Memory Models

It has been fifty plus years since James P. Egan (1958) wrote his famous technical report, Recognition Memory and Operating Characteristics, and yet recognition memory modelers are still debating over the correct theory of recognition memory. We believe all recognition models are wrong, but several of them have been very useful in such areas as medical science, ethnography, forensic psychology, memory and psychophysics. In particular, we believe the process underlying recognition memory is more complex than has historically been described given the few degrees of freedom available to a modeler. In this spirit, we take a new look at the Gaussian Signal Detection model (SDT) and the Double High-Threshold (2HT) models for old/new recognition memory data. Advocates of SDT do not like the restriction of equal variance for old and new items, and advocates of the 2HT model do not like the assumption that the detection probabilities for old and new items are equal. We add a third response option, Uncertain, and modify both models accordingly. The new models are saturated with

four parameters in a product trinomial data structure, but either one or both models cannot fit many possible data tables. To handle the sampling assumptions we develop Bayesian hierarchical versions of both models, and we fit both models to data allowing heterogeneity in both subjects and items. Our focus is not on proving one of the models is better than the other, but instead we focus on when both models give similar stories about sensitivity and bias.

2.1 Introduction

For the better part of the 20th century and start of the 21st century, many psychologists have demonstrated an uncanny devotion to the search of finding the *correct* scientific model of recognition memory by contending between two classes of models. The first and most influential is the Signal Detection Theory (SDT) model with origins tracing back to engineering in signal processing, and later adopted in psychology through its application to psychophysics of vision and audition by Swets, Tanner, & Birdsall (1955) and Green, Swets (1966). The second and often preferred model stems from threshold theory, with the most prominent example being the Double-High Threshold (2HT) model with origins tracing back to Blackwell's work on correcting for guessing, and later formalized by Snodgrass (1988).

From an early start, efforts aimed at dissociating the models have mainly revolved around each theory's prediction of the receiver operator characteristics (ROC; e.g. Egan, 1958). In fact, the perennial debate continues today (see Yonilinas, 2002, for review), using complex model comparison techniques for fitting the predicted ROC curves. While the precise shape of the ROC function is hotly debated, one persistent finding is that the shape of the ROC (whether rectilinear or curvelinear) depends more on the choice of task used by the experimenter than on a fundamental property of memory. In our view, some of this recent work has lost sight of the usefulness of both model families as measurement tools and thus

appears to us to be in service of selecting the scientifically correct model family with very little evidence to suggest this is statistically justifiable.

In a typical recognition memory experiment, N items are sequentially presented to a decision maker (DM) during a controlled study phase. After a predetermined delay, the DM is presented with a list containing previously studied items along with new distractor items. The presentation order is randomized to ensure an even distribution of old and new items. The DM is tasked to respond to each item on a test trial with a "yes" if the item was presented in the study phase, or "no" otherwise. To avoid missing data, participants are usually asked to provide their best guess before proceeding to the next item on the test phase. Variations of this simple Yes-No experimental design account for a majority of the comparisons between the two models.

The experimental design limits the DM to two response alternatives, forcing the DM to either guess at random or bias their response when they do not have sufficient evidence. Effects of forced guessing have been reported in the learning literature as exacerbating misinformation and partial information effects, thus reducing the unbiased memory characteristics measured by the models. To reduce the effects on the measurement of memory ability caused by forcing a DM to respond, we augment the set of possible response alternatives to include an additional response category (e.g. "Uncertain"). A consequence of adding another response category is to increase the degrees of freedom available for modeling which we utilize to further explore the two models.

In this paper we start by briefly comparing the simplest versions of the SDT model and the 2HT model and show that both are describing very similar properties of the data. Following this introduction we extend the data structure to allow a third option giving a DM the freedom to express their confidence. With this new data structure we introduce two new models, one under aegis of SDT and the other within the scope of 2HT theory. These two new extant models provide a chance for the exploration of additional parameters that both

SDT and 2HT theorists would like to see. We show that parameters in both models aim at describing similar mental processes similar to their basic model prototypes. Furthermore, we compare these models using a classical approach and a Bayesian framework and show that neither model is the *correct* model of recognition memory but rather both are useful statistical tools for specific data sets.

In the next section we extend the two choice recognition memory experiment to allow for an extra response. This addition increases the degrees of freedom (d.f.) to allow the estimation of two extra parameters for each class of models. The SDT model is extended to include a variance parameter for the signal trials and the 2HT is extended to include a different ability parameter for the noise distributions. Additionally, the data structure can be seen as a three-point confidence interval whose structure is similar to ROC data. Although the fit of ROC data is not the focus of this paper, the statistical structure of both extended models will be explored.

These models are designed for experimental situations where a participant is exposed to a list of words on a study trial, and then on a test trial the participant is exposed to a mix of old studied words and new distractor words. Usually words are presented one at a time and both the study words and test words are chosen to be semantically unrelated and drawn from some narrow range in such indices as word frequency and concreteness. For each presented test word, the subject says "Yes" if they think it is an old studied word, "No" if they think the test word is a new distractor, and "Uncertain" if they are unsure whether the word appeared in the studied list or not.

The goal of adding one additional response option is to extend the models in a simple nontrivial way to allow a precise study of the mathematical structures observed in more complicated versions of the two theories. Research on ROC curves obtained from recognition memory experiments have shown there to be an asymmetry not fully described by the simple two parameter versions of the models. Mainly the asymmetry found from studies suggest that

an additional process may be underlying memory in a recognition study that is independent of memory sensitivity already measured. To account for such irregularities researchers have suggested using an unequal-variances version of the standard SDT model along with an additional threshold for the high threshold models.

We begin by expanding the standard experimental conditions to include a third response category. The additional response category is sufficient to allow the estimation of two more parameters for each model. We will examine the effect of the new models on the well-established experimental word frequency effect (WFE), which is a counterintuitive effect resulting from varying linguistic frequency conditions, in order to gain additional insight into the latent memory processes involved by using these more complete quantitative models.

2.2 Ternary response data

Consider the data for a single group, ternary-response recognition memory experiment, where M participants are asked to study a set of N_1 items during a study phase. Afterwards, each participant is tested on the old set of N_1 items along with $N_2 = N - N_1$ distractor items. On the test trial, three response options are available to the participant to choose from. The first two are the basic yes-no responses, and the last is an "uncertain" response. Formally, the two independent trichotomous random variables, X_1 for a response to an old test item and X_2 for a response to a new test item are defined as:

$$X_i = \begin{cases} 1 & \text{if "Yes"} \\ 0 & \text{if "Uncertain"} \\ -1 & \text{if "No"} \end{cases} \tag{2.1}$$

For conciseness, we will define the marginal probability functions as $p_{o,1}$ to indicate $Pr(X_{ik} = 1|Signal)$, where the o is for old stimulus and the number 1 corresponds to the response a subject indicated. By continuing this notation, we can see that $Pr(X_{ik} = 1|Noise)$ can be written as $p_{n,1}$ and the notation for the rest of the conditional probabilities can be seen in Table 1.

Table 2.1: Conditional Probabilities

	"Old"	"Uncertain"	"New"
Signal	$p_{o,1}$	$p_{o,2}$	$p_{o,3}$
Noise	$p_{n,1}$	$p_{n,2}$	$p_{n,3}$

Any statistical model developed for the two trichotomous random variables will inevitably induce a partition on the probability space given a particular model's parameters specifications. In other words, parametric models specify a partition on the probability space,

$$\pi = \{(p_{o,1}, p_{o,3}, p_{n,1}, p_{n,3}) : 0 < p_{o,1} < 1, 0 < p_{o,3} < 1, 0 < p_{n,1} < 1, 0 < p_{n,3} < 1\}. \quad (2.2)$$

where Π is the space of all possible marginal probabilities in a ternary-response recognition memory experiment. While the chosen probabilities do provide the required information, it is important to note that any other pair (e.g. $\{p_{o,2}, p_{o,3}\}$ and $\{p_{n,2}, p_{n,3}\}$) for a given random variable is equally informative since they all sum to 1.

2.3 Signal Detection Model

The first proposed model is a version of the unequal variance SDT (UVSDT), updated to reflect a third response category of "uncertain" (referred to hereafter as 3R-SDT). Evidence

from multiple studies has encouraged the use of unequal variance by fixing the variance, σ_o, of the signal distribution to 1.25. For our purposes, we allow σ_o to be free to vary in $[1, \infty)$ such that that it allows the standard deviation of the signal distribution to be different than the standard deviation of the noise distribution. We begin by making the assumption that familiarity of each item presented during the test trial is a draw from one of two Gaussian distributions on the continuum with means $\{\mu_o, \mu_s\}$ and variances $\{\sigma_o, \sigma_n\}$ corresponding to each class of items. A decision for one of the three response categories is determined by two partitions to the familiarity continuum.

The mutually exclusive and exhaustive partitions to the familiarity axis become apparent with the two decision criteria such that items above the first threshold, τ_1, are judged old, and items that are below a second threshold, τ_2 are judged new. The area between these two decision criteria represents a region of uncertainty; where items within this critical interval fail to elicit sufficient evidence for a decisive judgment. Though 3R-SDT does not focus on the interval between the two criteria, it has been previously studied for remember-know judgments (Donaldson, 1996).

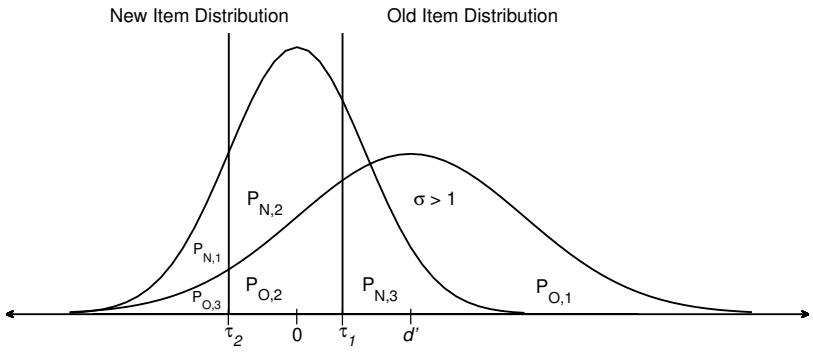

Figure 2.1: 3R- Signal Detection Theory Model. Here the "new" stimuli distribution represents noise.

As it stands, the model exemplifies the theoretical assumptions of SDT thus permitting one to draw conclusions of psychological variables using the parameters. With the additional response we are able to relax the equal variance assumption usually instantiated for two

response recognition memory caused by a lack of degrees of freedom. Figure 2.1 displays the model's two Gaussian distributions along with the two criteria. It is important to point out the σ located within the signal distribution. The parameter σ is the standard deviation of the old stimuli distribution. Since the theory assumes that each signal observation contributes to the variance of the signal density function, σ must be greater than 1 (Wixted, 1992).

Once again parametric models define a particular set of probability distributions over the two random variables in Equation 1 for each possible parameter combination. The SDT model is specified by parameters $\Theta = (d', \sigma, \tau_1, \tau_2)$ where $\Omega_\Theta = \{(d', \sigma, \tau_1, \tau_2) : d' \in [0, \infty), \sigma \in [1, \infty), \tau_1 \in (-\infty, \infty), \tau_2 \in (-\infty, \infty)\}$. An important thing to note here is that when the SDT model is extended to a UVSDT model, the assumption of a curvilinear receiver operator characteristic (ROC) curve is no longer valid (Swets, & Tanner, & Birdsall (1961)).

2.4 Basic High Threshold Model

The second model proposed in this paper is based on the double high threshold theory. As seen in Figure 4, the subject either detects the signal observation (with probability D), or does not (with probability $1 - D$). When a subject does not detect the observation as belonging to the signal state structure, then the subject makes a decision of whether or not to guess. The subject chooses to guess with probability α, and if the subject chooses not to guess (with probability $1-\alpha$), then the subject will respond with "uncertain". If the subject chooses to guess then the response would be either "yes" (with probability γ), or "no" (with probability $1 - \gamma$). A pictorial representation of this new model can be seen in Figure 4.

This image also shows that a subject uses a new threshold, B, for noise observations. When a noise trial is presented to a subject, the model would allow the subject to discriminate that occurrence from the other signal or noise trials stored in memory, (with probability B).

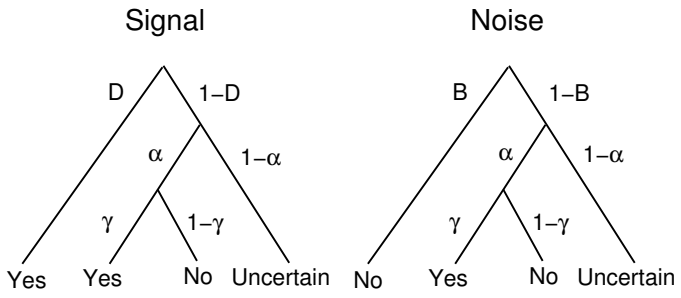

Figure 2.2: New Double High Threshold Model

If the subject does not discriminate the new item as new, then with the probability α the subject would guess or with probability $1 - \alpha$ the subject may choose not guess.

Using only one α and one γ for both state trees assumes that their guessing threshold is the same when they are either in the detect signal state or discriminate noise state. Essentially, when a subject is not able to correctly detect a signal observation or correctly discriminate a noise observation then the subject will revert back to the same process of guessing.

The new 3R-2HT model is specified by parameters $\Delta = \{D, B, \gamma, \alpha\}$, with parameter space $\Omega_\Delta = \{(D, B, \gamma\, \alpha) : D \in [0,1], B \in [0,1], \gamma \in [0,1], \alpha \in [0,1]\}$. Once again the direction of this paper is not to introduce new theories of recognition memory, nor to create new models with the goal of finding the correct model of recognition. Rather, the aim is to develop new measurement tools for data that may include an unknown response.

2.5 Comparing both new models

In this section we compare the two new models with three response options designed for an episodic memory task. The comparison is not intended to support a particular model,

rather, we aim to describe the two models statistically.

The problem of calculating the closed form expressions remains tractable with a product trinomial likelihood function. For the sake of brevity, the likelihood functions are here omitted for these models; however, the reader who wishes to work with the analytical forms may derive the likelihood functions by substituting terms introduced in Observations 3-4. The closed form expressions for the four 3R-SDT model parameters are:

$$\hat{d'} = \frac{\Phi^{-1}(p_{o,1})\Phi^{-1}(p_{n,3}) - \Phi^{-1}(p_{n,1})\Phi^{-1}(p_{o,3})}{\Phi^{-1}(p_{o,1}) + \Phi^{-1}(p_{o,3})} \tag{2.3}$$

$$\hat{\sigma} = \frac{\Phi^{-1}(p_{n,1}) + \Phi^{-1}(p_{n,3})}{\Phi^{-1}(p_{o,1}) + \Phi^{-1}(p_{o,3})} \tag{2.4}$$

$$\hat{\tau_1} = \frac{-2\Phi^{-1}(p_{o,1})\Phi^{-1}(p_{n,1}) - \Phi^{-1}(p_{o,1})\Phi^{-1}(p_{n,3}) - \Phi^{-1}(p_{o,3})\Phi^{-1}(p_{n,1})}{2[\Phi^{-1}(p_{o,1}) + \Phi^{-1}(p_{o,3})]} \tag{2.5}$$

$$\hat{\tau_2} = \frac{\Phi^{-1}(p_{o,1})\Phi^{-1}(p_{n,3}) - \Phi^{-1}(p_{o,3})\Phi^{-1}((p_{n,1}) + 2\Phi^{-1}(p_{n,3}))}{2[\Phi^{-1}(p_{o,1}) + \Phi^{-1}(p_{o,3})]} \tag{2.6}$$

where, $\hat{d'} > 0$, $\hat{\sigma} \geq 1$, $\hat{\tau_2} < \hat{\tau_1}$ and $\Phi(.)$ is the quantile function of the Gaussian cumulative distribution function. We can see that as $P_{o,1} \to 1$ and $P_{n,3} \to 1$, as in the case where discriminability is increased to obtain near perfect judgments, $d' \to \infty$ and $\sigma_o \to 1$.

The MLE's for the new 3R-2HT model parameters are:

$$\hat{D} = \frac{p_{o,1}p_{n,2} - p_{o,2}p_{n,1}}{p_{n,2}} \tag{2.7}$$

$$\hat{B} = \frac{p_{o,2}p_{n,3} - p_{o,3}p_{n,2}}{p_{o,2}} \tag{2.8}$$

$$\hat{\gamma} = \frac{p_{o,2}p_{n,1}}{p_{o,3}p_{n,2} + p_{o,2}p_{n,1}} \tag{2.9}$$

$$\hat{\alpha} = \frac{p_{o,3}p_{n,2} + p_{o,2}p_{n,1}}{p_{o,2}p_{n,2} + p_{o,3}p_{n,2} + p_{o,2}p_{n,1}} \tag{2.10}$$

where $0 < (\hat{D}, \hat{B}, \hat{\gamma}, \hat{\alpha}) < 1$

Again, constraints on the product trinomial space of the data are imposed out of concern that the parameters in each model remain within their respective distributions. Formally:

Observation 3. $\Phi^{-1}(p_{o,1}) + \Phi^{-1}(p_{o,3}) > \Phi^{-1}(p_{n,1}) + \Phi^{-1}(p_{n,3})$ and $\Phi^{-1}(p_{n,1}) + \Phi^{-1}(p_{o,3}) > \Phi^{-1}(p_{o,1}) + \Phi^{-1}(p_{n,3})$ if and only if there is a unique $\{d', \sigma, \tau_1, \tau_2\} \in \Omega_{3R-SDT}$ with

$$\begin{array}{ll} P_{o,1} = \Phi(\frac{\frac{d'}{2}-\tau_1}{\sigma}) & P_{n,1} = \Phi(-\frac{d'}{2}-\tau_1) \\ P_{o,2} = \Phi(\frac{\frac{d'}{2}-\tau_2}{\sigma}) - \Phi(\frac{\frac{d'}{2}-\tau_1}{\sigma}) & P_{n,2} = \Phi(-\frac{d'}{2}-\tau_2) - \Phi(-\frac{d'}{2}-\tau_1) \\ P_{o,3} = \Phi(\frac{\tau_2-\frac{d'}{2}}{\sigma}) & P_{n,3} = \Phi(\frac{d'}{2}+\tau_2) \end{array}$$

Observation 4. $p_{o,1} > \frac{p_{o,2}p_{n,1}}{p_{n,2}}$ and $p_{n,3} > \frac{p_{o,3}p_{n,2}}{p_{0,2}}$ if and only if there is a unique $\{D, B, \alpha, \gamma\} \in \Omega_{3R-2HT}$ with

$$\begin{array}{ll} P_{o,1} = D + (1-D)\alpha\gamma & P_{n,1} = (1-B)\alpha\gamma \\ P_{o,2} = (1-D)(1-\alpha) & P_{n,2} = (1-B)(1-\alpha) \\ P_{o,3} = (1-D)\alpha(1-\gamma) & P_{n,3} = B + (1-B)(1-\gamma)\alpha \end{array}$$

Clearly the demarcations do not partition the sample space of the data equally for both models. Generally the model with the larger prediction space is considered too flexible and in model selection techniques such as the normalized maximum likelihood, it is penalized. The shared constraints permitted each model the same a priori probability of fitting a pair $\{P_o, P_n\}$ in the sample space of the data. In return, the correspondence between the model parameters proved to be essential in showing them to be statistically equivalent. Now, the current models require stricter boundary conditions on Ω_{Data}, which may yield different a priori probabilities of fit for each model. In fact, we can numerically check the polarity between the two models on Ω_{Data}. To see this, a full grid is used to find every combination of the product trinomial distribution for different sample sizes. Then, the boundary conditions imposed on the data space for each model are used to find the region in Ω_{Data} that can be fit by each model. Once again, we find every combination of the product trinomial distribution using four experimentally convenient sample sizes.

Table 2.3 shows that 3R-SDT can evaluate a larger number of possible data patterns than 3R-2HT. However, a serious problem that is often overlooked is what Green and Swets (1966)

Table 2.2: The sampled regions pertaining to one, both, or neither model.

N	3R-SDT Only	3R-2HT Only	Both	Neither
25	14,270 (11.58%)	10,558 (8.57%)	8,462 (6.87%)	89,911(72.98%)
50	247,315 (14.07%)	153,697 (8.74%)	132,991 (7.56%)	1,224,273 (69.63%)
100	4,071,648 (15.35%)	2,288,685 (8.63%)	2,106,780 (7.94%)	18,065,688 (68.09%)
200	66,000,188 (16.01%)	35,109,704 (8.52%)	33,467,041 (8.12%)	277,553,668 (67.35%)

called the knotty theoretical problem in SDT. A lack of monotonicity in the likelihood ratio can lead to higher FA rates than H rates and lower CR than M rates. This translates to having more area under the signal distribution to the left of the criterion when compared to the same region of the noise distribution and is prone to occur when $\sigma_o > 1$. In fact, it is easy to show that the point, z^*, at which the rate of saying 'new' to old items is greater than the rate of saying 'new' to new items is approached rapidly. Equating the area to the left of an arbitrary point z^* for both the new and old distribution we get, $\frac{z^* - \frac{d'}{2}}{\sigma_o} = z^* + \frac{d'}{2}$ and solving for z^* we get:

$$z^* = \frac{d'}{2}\frac{1+\sigma_o}{1-\sigma_o} \tag{2.11}$$

Solving (11) when $d' = 1$ and $\sigma_o = 2$ we see that $z^* = -1.5$ which is only 1 standard deviation away from the mean of the noise distribution. An observation made by Roberts and Pashler (2000) and evident in our results is that overly flexible models that are able to fit many different data sets tend to include data sets seen as inconsistent with core assumptions. So to discard these inconsistencies we reanalyze the complete sample space, Ω_{Data}, with the additional constraint $P_{o1} > P_{n1}$ and $P_{o3} < P_{n3}$.

The permissible number of data patterns available from the complete set is reduced to one third of its original size. No change in the area of best fit was detected for the 3R-2HT

and a reduction of about one half to the 3R-SDT's measurable space. While the models are not quite the same in their number of best fitted data patterns, there still remains the question of which data patterns are best fitted by one model alone. Since this question may be answered through psychological experiments, we will return to it in the next section.

2.5.1 Correlation

With the information provided so far, it is of interest to ascertain whether the parameters in these new models correlate like those in the basic models. In order to do this we found the constraints on the data for each model. Without the use of constraints on the data, the estimated parameters $\hat{D}, \hat{B}, \hat{d'}$, and $\hat{\sigma}$ would not be within the correct range. The theory of signal detection assumes that the parameter $\hat{d'}$ has to be greater than 0 otherwise there would be no information available on a subject's discrimination index. The theory also assumes that each signal observation contributes to the variance of signal distribution. Over time this accumulation of variance should make the parameter σ greater than the standard deviation of the noise distribution, which is commonly set at 1. To account for these two assumptions the following equations are the constraints needed:

Table 2.3: Correlation for expanded models

	$\hat{D}$	$\hat{B}$	$\hat{\gamma}$	$\hat{\alpha}$	$\hat{d'}$	$\hat{\sigma}$	$\hat{\tau_1}$	$\hat{\tau_2}$	$\hat{\tau_1} - \hat{\tau_2}$
$\hat{D}$	1								
$\hat{B}$	.1543	1							
$\hat{\gamma}$	.2225	-.3719	1						
$\hat{\alpha}$	.0132	-.0705	-.0410	1					
$\hat{d'}$	.7823	.3013	-.1923	-.0954	1				
$\hat{\sigma}$	.4218	.0442	-.4543	.1768	.7053	1			
$\hat{\tau_1}$	-.1176	.4482	-.7975	-.4113	.3056	.3849	1		
$\hat{\tau_2}$	.0118	.7554	-.6576	.5291	.2073	.3211	.4287	1	
$\hat{\tau_1} - \hat{\tau_2}$	-.1246	-.2044	-.1824	-.8740	.1100	.0847	.5842	-.4827	1

The area between the two thresholds mentioned above, shown as $\tau_1 - \tau_2$, is also used in

the correlation test. The reason for this is that this area of uncertainty is analogous to the probability that the subject enters the guess state or not in the threshold model because in both cases there is not enough information to make a correct judgment. This interpretation is supported by the strong correlation found in Table 2.4.

It is interesting to note that the correlation found between D and d' is strong but not as strong as the basic model comparison. The drop in correlation can be attributed to the inclusion of the parameter B because the latent measure D represented both the instance when old items were correctly identified as old and when new items were correctly identified as new. Now, this conglomeration of latent abilities has been separated, thus reducing the relationship with the SDT model's discrimination parameter. Another aspect of this table that is interesting is the correlation between the B parameter and τ_2; by referring back to Figure 2.1 one can note that the τ_2 cut-point is located closest to the noise distribution. Thus whenever a subject passes the uncertainty threshold towards correctly identifying a noise stimuli, this is similar to the B parameter that gives the probability a subject is able to discriminate the noise stimuli from the rest of the signal stimuli.

2.5.2 Hierarchical Model Framework

A hierarchical modeling framework is adopted to allow for individual differences among the parameters of each model. The hierarchical framework provides an intuitive method of including person-specific latent variables by defining population distributions on the parameters. The choice of population distributions reflects important theoretical assumptions made by each model, and must be consistent with the range set forth by model specifications. We begin by defining the population distributions and their hyperparameters for the 3R-SDT followed by the 3R-2HT.

For the new 3R-SDT, the Gamma distribution function is assigned as the population distri-

bution for the memory sensitivity parameter, $d'_i \in [0, \infty)$. The gamma distribution function is parameterized by a shape parameter $\alpha_{d'}$ and rate parameter, (i.e inverse scale parameter) $\beta_{d'}$ both of which are expressed on $\mathbb{R}^+$ and are not subject dependent. The population probability density function for d'_i is:

$$f(d'_i, \alpha_{d'}, \beta_{d'}) = \frac{\beta_{d'}^{\alpha_{d'}}}{\Gamma(\alpha_{d'})} d_i'^{\alpha_{d'}-1} e^{-\beta_{d'} d'_i}$$

where Γ is the gamma function i.e. $\Gamma(n) = (n-1)!$ and e is Euler's number.

For our purpose, characteristics of the population such as the mean, $\mu_{d'}$ and variance $\sigma^2_{d'}$ are more meaningful. These statistics are functions of $\alpha_{d'}$ and $\beta_{d'}$ such that:

$$\mu_{d'} = \frac{\alpha_{d'}}{\beta_{d'}}$$

$$\sigma^2_{d'} = \frac{\alpha_{d'}}{\beta^2_{d'}}$$

By rearranging the terms and substituting them into the gamma distribution function, we arrive at:

$$d'_i \sim \text{Gamma}(\frac{(\mu_{d'})^2}{\sigma^2_{d'}}, \frac{\mu_{d'}}{\sigma^2_{d'}}) \tag{2.12}$$

Given that the inverse-gamma distribution is known to be the conjugate prior of the variance parameter of a normal distribution, we assign the signal distribution's standard deviation,

$\sigma_{oi} \in [1, \infty)$, an inverse-gamma distribution function.

$$f(\sigma_i'; \alpha_\sigma, \beta_\sigma) = \frac{\beta_\sigma^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)} \sigma_i^{-\alpha_\sigma - 1} e^{\frac{-\beta_\sigma}{\sigma_i}} + 1$$

The function is shifted to the right by one to satisfy model specifications posed by the theory of 3R-SDT. Again, the mean and variance of the distribution is calculated and solved for the inverse gamma distribution parameters:

$$\mu_{\sigma_o} = \frac{\beta_\sigma}{\alpha_\sigma}$$

$$var_{\sigma_o} = \frac{\beta_\sigma^2}{(\alpha_\sigma - 1)^2(\alpha_\sigma - 2)}$$

$$\sigma_{oi} \sim \text{Inverse-Gamma}(\frac{(\mu_\sigma)^2}{\sigma_\sigma^2}, \frac{\mu_\sigma}{\sigma_\sigma^2}) \tag{2.13}$$

To include person specific model parameters for the 3R-SDT model's two threshold parameters, we must consider the range over the $\mathbb{R}$ permissible for each criterion. Since the theory does not restrict the thresholds to a closed interval, the values are drawn from two Gaussian

distributions with $\{\mu_{\tau_1}, \mu_{\tau_2}\}$ and $\{\sigma_{\tau_1}, \sigma_{\tau_2}\}$.

$$\tau_{1i} \sim \text{Gaussian}(\mu_{\tau_1}, \sigma_{\tau_1})\text{T}(\tau_2, \infty) \tag{2.14}$$

$$\tau_{2i} \sim \text{Gaussian}(\mu_{\tau_2}, \sigma_{\tau_2}) \tag{2.15}$$

Model specifications restrict the values of $\tau_1 > \tau_2$ so T is used to truncate τ_1's distribution to values greater than τ_2.

A long-standing property of threshold models is the unit interval distribution on the parameters. The beta distribution function is a reasonable choice so long as the desired analysis does not include correlation between the parameters. However, a simplifying assumption has been to coalesce D_i and B_i into a single ability parameter without loss of generality. Here we test this by means of Pearson's r correlation to show the strength of a linear relationship between the two parameters. To remove the need for a two step approach of estimating the parameters followed by correlating them to each other, the hierarchical approach provides a natural method of involving covariate information in the model. In order to avoid problems of overestimation caused by using the beta probability distribution (Oravatz, Anders, Batchelder, 2013), we use the bivariate Gaussian probability distribution function.

This is possible since the integral transform theorem proves that any continuous normal probability distribution with mean and variance, $\{\mu, \sigma^2\}$, respectively, can be converted to a uniform distribution in [0,1]. This is usually done with a link function such as the logit or probit. To remain consistent, we chose the quantile function associated with the Gaussian

probability distribution function.

$$\begin{bmatrix} \Phi^{-1}(D_i) \\ \Phi^{-1}(B_i) \end{bmatrix} \sim \text{Bivariate-Normal}(\mu_{(D,B)}, \Sigma_{(D,B)}) \tag{2.16}$$

where $\Sigma_{(D_i,B_i)}$ is the covariance matrix of person-specific ability parameters (D_i, B_i) and $\Phi^{-1}(\cdot)$ is the inverse cumulative Gaussian distribution function, i.e., the probit function.

The remaining parameters of the new 2HT are not assumed to covary, so the choice for their probability distributions functions is simply:

$$\Phi^{-1}(\gamma_i) \sim \text{Normal}(\mu_\gamma, \sigma_\gamma) \tag{2.17}$$

$$\Phi^{-1}(\alpha_i) \sim \text{Normal}(\mu_\alpha, \sigma_\alpha) \tag{2.18}$$

where once again $\Phi^{-1}(\cdot)$ is the quantile function of the cumulative Gaussian distribution function (i.e. the probit function).

2.6 Experiment

The discussion so far has focused on describing the new models and their relationship to each other. To relate these model predictions to psychology we collected data from two

experiments. A well-documented result of recognition memory testing is the counterintuitive mirror effect with varying linguistic frequency conditions. The mirror effect refers to results obtained from varying extra-experimental conditions such as word frequency, where the hit rate increases and the false alarm rate decreases due to a change in one condition.

While the effect has been noted using other experimental manipulations (Glanzer & Glanzer; Stretch & Wixted, 1998), we will focus on results obtained from a change in linguistic frequency and refer to it as the word frequency effect (WFE). The WFE is characterized by better performance of low linguistic frequency (LF) words than high linguistic frequency (HF) words (Gorman, 1961; Shepard, 1967). This result has been observed so frequently it has been described as an empirical regularity of recognition memory (Glanzer, Adams, Iverson, & Kim, 1993).

Many theoretical explanations have been proposed for the WFE. For example, Shiffrin & Styevers, (1997) proposed that HF words may be made up of less distinctive lexical features which presumably adversely affects the memorability of a word. This assumption was tested by Malberg et al. 2002, when they examined the responses of a yes-no recognition memory test designed to vary a low order measure of orthographic features. Malberg et al reported that words composed of less frequently used letters were better recalled than words containing more frequently used, letters as predicted by Shiffrin & Styevers (1997). However, Malberg et al. also noted that when orthographic features were controlled in LF and HF words, the mirror effect was still evident, suggesting that orthographic features may not sufficiently account for the phenomenon.

Our interest in WFE is not to contribute to the ever expanding list of theoretical explanations, rather it is to attain some insight into latent memory processes underlying the WFE by using more complete quantitative models. We begin by expanding the standard experimental conditions to include a third response category. The additional response category is sufficient to allow the estimation of two more parameters for each model, as previously

described.

2.6.1 Subjects

Forty-nine college-aged undergraduates from the University of California, Irvine were recruited from the university's psychology department subject pool. Subjects received course credit in exchange for their participation in the experiment. All 49 subjects participated in two experiments 1.A and 1.B.

2.6.2 Design

The study consisted of a within subject design where each subject studied, for later testing, both low frequency words and high frequency words. Neither the researcher nor the subject were made aware of which frequency list would be presented first. For each subject a random subset of words from each list was chosen to be used for the study list. No two subjects studied the same list of words.

2.6.3 Stimuli

Two word lists (low and high frequency word lists) were taken from the MRC psycholinguistic database (Coltheart 1981). Each list consisted of 80 nouns with each word containing five to eight letters. The average written frequency of the lists was 4.125 and 36.425 for the low frequency and high frequency word lists, respectively (Kucera and Francis 1967). Forty words from each list were randomly sampled and assigned to each subject without replacement for the studied list and the remaining 40 were assigned as lures for the same cycle.

2.6.4 Procedure

At the beginning of the experiment each subject was instructed to sit in front of a computer monitor approximately 2 ft away. At the start of the experiment subjects were told that during the study phase a single word would appear in the middle of the screen. The participant had as much time as they needed with each word. In order to move onto the next word, every subject was told to press the space bar. As a precaution, the program was designed to lock the keyboard keys for one second after the presentation of the word. This was the same for all 40 words in both study phases.

On the test phase, subjects were presented with 80 words. Each randomly chosen word was situated in the middle of the screen until the subject pressed a key on the keyboard. Each subject was asked to indicate whether the word being presented belonged to the study list by pressing the V key on the keyboard. If the word was not previously seen, they were asked to press the N key. Finally, the subjects were informed that if they were uncertain a word was previously presented, they where allowed to indicate "don't know" by pressing the B key on the keyboard. A notecard was placed in front of each subject with these key codes for response types. The test phase did not have the 1 second restriction placed on the keys in the study phase. After finishing the first study-test phase, subjects took 5-minute break before returning back for the second recognition test.

2.6.5 Bayesian Estimation Inference

The data is analyzed using the hierarchical models with a Bayesian Estimation inference approach. In the last decade it has become an increasingly popular approach to estimating hierarchical cognitive models, e.g. Lee & Wagenmakers (2014). The contemporary approach is adopted to facilitate the estimation of model parameters from the hierarchical framework. While classical approaches such as maximum likelihood exist, the statistical inference

techniques do not provide a straight forward method of estimating the parameters of a hierarchical model. In fact, as noted earlier, not all integrals over the data have closed-form solutions so the results would require the use of finite sums.

The advantage of using Bayesian Estimation techniques is the avoidance of high-dimensional integration over many random-effect distributions. Furthermore, information about the posterior distributions of the parameters is readily available thus adding a greater degree of confidence by providing credible intervals. Careful consideration for the choice of each parameter distribution was outlined earlier, using differing boundary conditions on the sample space to avoid biasing model comparison measures in favor of one over the other.

The hierarchical structure becomes transparent with the designation of hyperprior-distributions for each model parameter's distribution. We begin once again by defining the hyperpriors and their distributions for the 3R-SDT model followed by the 3R-2HT model.

The following hyperprior distributions correspond the the 3R-SDT parameters[1]:

$$\{\mu_d, \mu_\sigma, \mu_{\tau_1}, \mu\tau_2\} \sim \text{Normal}(0, 1)$$

$$\{\sigma_d, \sigma_\sigma, \sigma_{\tau_1}, \sigma_{\tau_2}\} \sim \text{Inverse-Gamma}(1, 1)$$

The inverse Gamma function is the conjugate prior for variance parameters so we assign the Gamma function to the variance parameter. The hyperprior distribution is uninformative in the sense that no a priori information is imposed on the parameters (Gelman 2004).

The bivariate normal population distribution is parameterized by the mean vector $\mu_{(D,B)}$ and the covariance matrix $\Sigma_{(D,B)}$. We fix the $\mu_{(D,B)}$ vector at $([0, 0])$ and the covariance

[1]The priors are not indexed by subject or item since the person-specific parameters are draws from the population distribution.

matrix is modeled using the inverse-Wishart function.

$$\Sigma_{(D,B)} \sim \text{Inverse-Wishart}(\mathbf{I},\text{df})$$

where $\mathbf{I}$ is the identity matrix.

The remaining hyperpriors are normally distributed with mean zero and variance 1. The model code used in JAGS for both models is in the appendix.

2.7 Results

The average response proportions obtained for the LF and HF conditions are presented in Table 6. Along with response proportions, the p-values of t-tests are presented on the far right column. The t-test was conducted without controlling for multiple comparisons. The response proportions show the expected patterns for the mirror effect: the probability of correctly identifying an old word for the LF condition is higher than for the HF condition and the probability of incorrectly identifying a new HF word is higher than for the LF condition.

Table 2.4: Average Response Proportions for LF and HF Words (SD).

	Low Frequency	High Frequency	p
$p_{o,1}$	0.7574 (0.1344)	0.7144 (0.1445)	0.0100
$p_{o,2}$	0.0761 (0.0894)	0.0830 (0.0780)	0.4269
$p_{o,3}$	0.1665 (0.1260)	0.2027 (0.1464)	0.0199
$p_{n,1}$	0.0936 (0.0914)	0.1106 (0.0923)	0.2027
$p_{n,2}$	0.1277 (0.1448)	0.1537 (0.1573)	0.1553
$p_{n,3}$	0.7787 (0.1696)	0.7356 (0.1774)	0.0658

A test of individual differences was conducted to confirm the move towards using a hierarchical version of the model. It is standard practice to aggregate subject responses if the assumption of homogeneity is met so to check this, we use a permutation test for individual differences (Smith & Batchelder, 2008). The chi-square test of independence for both data sets are: χ^2(df=48,.01) and χ^2(df=48,.01). The results indicate strong evidence against aggregating participant responses.

2.8 Conclusions

Recognition memory experiments based on varying degrees of linguistic frequency often show a paradoxical finding that HF words are less likely to be recognized than LF words. This finding was replicated in our experiments using three response categories rather than the usual two. Although the difference in misidentification of new words was not significant, the data exhibited the quintessential mirror effect pattern. The purpose for the experiment was to allow the updated versions of both theories to provide insight into the latent processes occurring for both low and high frequency lists and to showcase the similarities between them. A finding shared by both models is that the mirror effect is not a function of individual bias but rather a function of the memory sensitivity.

The 3R-2HT model demonstrated that a change in linguistic frequency did not greatly alter a subject's bias to respond old or new. Thus any bias held by a subject on the first experimental condition remained the same on the second condition. Now, in the study by McCormack & Swenson (1972), SDT was used to determine if their data agreed to normality and homogeneity properties of SDT. In their analysis, they found that for both linguistic frequency conditions the signal distribution was more variable than the noise distribution. Although they relied on the model with two (rather than three) parameters, they were able to find this result by checking the slopes of the memory operator characteristics. In our

current version of 3R-SDT, we also find that the two conditions are best represented with different variances on the signal distributions.

Using a signal detection theory (SDT) model, Stretch & Wixted (1997) examined through a series of five experiments the hypothesis that a change in criterion is responsible for the differences in response characteristics between LF and HF, and determined this hypothesis was not supported. Stretch and Wixted concluded that an increase in incorrect recognition of new HF words might be caused by a higher sense of familiarity as predicted by Glanzer & Bowles (1976). A look at Stretch & Wixted's experimental procedure reveals that their design involved presenting the two different list of words at the same time in both the study and test phase. It is possible that by combining the two lists, a participant may be using the same fixed criterion. In order to account for this possibility, the current experiment presents the lists separately to participants.

Whether a person's ability of detecting a signal is caused by interference of highly associated words, such as those in the high frequency list, cannot be extrapolated from D_o. The reason for this is that the memory traces stored during the study are either interfering with each other, or the scope of common words embedded in our memory itself is causing interference. If the words stored during the study trial are interfering with each other's chance of being detected then the discrimination of new words should not be affected by word frequency. If, however, HF words share higher associations with other common words then the discrimination of new words should show an effect contingent on linguistic frequency. The results show a lower ability to discriminate new HF words. This suggests that interference is not based on the commonality between HF words in the list but rather the relationships rooted in our memory for common words.

Furthermore, the standard deviation parameter in SDT does not readily provide information about the latent processes without assumption, but paired with the d' parameter, it can show a more succinct story. The signal distribution pertaining to low linguistic frequency words

is centered on the mean much closer than the signal distribution for the high linguistic frequency list. The larger variance in the signal distribution for the FH list and smaller d' value suggests that it is harder to differentiate the signal from the noise for high frequency words compared to words with low frequency. Note that this difficulty does not mean that the strength of the memory trace is diminished with increased linguistic frequency. Instead, SDT limits the comparison to relative strengths of the signal distribution to the noise distribution across conditions.

A common feature of these proposals is that familiarity of HF words negatively impacts recognition performance. It is conceivable that a decrease in recognition performance for common words depends on a decision process adopted by our memory system. An immediate consequence of this notion is the possibility that a decision for LF and HF words depend on two distinct decision criteria.

Although it is not feasible to measure the level of familiarity each word has in episodic memory, it possible to analyze recognition performance using two quantitative models often used in conjunction with episodic memory experiments. By using an augmented 2HT model, we can test the assumption that performance on foil HF words, independent of performance of old HF words, is influenced by greater familiarity of HF words. Support would be shown if a greater reduction in the independent recognition performance of new words were to occur for HF words compared to LF words. As mentioned above, the SDT model has been used to test the hypothesis that a change in criterion is responsible for the mirror effect. While the hypothesis was not supported, it is possible that a decrease in performance for old HF words is a result of greater episodic familiarity of old HF words. With an augmented SDT model, we can test the assumption that a broader range of familiarity strengths negatively influences recognition of old HF words. If so, the results should show a more diffuse distribution of old HF words in addition to a smaller discrimination index, rather than a criterion shift.

2.9 References

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). Psychological Bulletin, 139, 1204?1212. 10.1037/a0033894

Bamber, D., & van Santen, J. P. (2000). How to assess a model's testability and identifiability. Journal of Mathematical Psychology, 44(1), 20-40.

Egan, J. P. (1958). Recognition memory and the operating characteristic. USAF Operational Applications Laboratory Technical Note.

Glanzer, M. A. JK (1985). The mirror effect in recognition memory. Memory & Cognition, 13(1), 8-20.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. Psychological review, 100(3), 546.

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. Journal of experimental psychology, 61(1), 23.

Swets, Tanner & Birdsall *The evidence for a decision making theory of visual detection*. Technical Report No. 40, Electronic Defense Group, University of Michigan, Ann Arbor, 1955.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. 1966. New York, 888, 889.

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory. A user?s guide.

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. Psychological Bulletin, 139(6), 1173.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological review, 107(2), 358.

Sikstrom, S. (2001). The variance theory of the mirror effect in recognition memory. Psychonomic Bulletin & Review, 8(3), 408-438.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. Journal of verbal Learning and verbal Behavior, 6(1), 156-163.

Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. Psychonomic Bulletin & Review, 4, 145-166.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. Journal of Experimental Psychology: General, 117(1), 34.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. Journal of memory and language, 46(3), 441-517.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. Psychological review, 114(1), 152.

Oravecz, Z., Anders, R., & Batchelder, W. H. (2015). Hierarchical bayesian modeling for test theory without an answer key. Psychometrika, 80(2), 341-364.

Chapter 3

A Cognitive Psychometric Model for the Psychodiagnostic Assessment of Memory-Related Deficits

Clinical tests used for psychodiagnostic purposes, such as the well-known Alzheimer's Disease Assessment Scale, cognitive sub-scale (ADAS-Cog), include a free recall task. The free recall task taps into latent cognitive processes associated with learning and memory components of human cognition, any of which might be impaired with the progression of Alzheimer's disease. A Hidden Markov Model of free recall is developed to measure latent cognitive processes used during the free recall task. In return these cognitive measurements give us insight into the degree to which normal cognitive functions are differentially impaired by medical conditions such as Alzheimers disease and related disorders. The model is used to analyze the free recall data obtained from healthy elderly participants, participants diagnosed as having mild cognitive impairment, and participants diagnosed with early AD. The model is specified hierarchically to handle item differences due to the serial position curve in free recall as well as within group individual differences in participants recall abilities. Bayesian hierarchical

inference is used to estimate the model. The model analysis suggests that the impaired patients have: 1) long-term memory encoding deficits; 2) short-term memory retrieval deficits for all but very short time intervals; 3) poorer transfer into long-term memory for items successfully retrieved from short-term memory; and 4) poorer retention of items encoded into long-term memory after longer delays. Yet, impaired patients appear to have no deficit in immediate recall of encoded words in long-term memory or for very short time intervals in short term memory.

3.1 Introduction

The most notable symptoms associated with Alzheimer disease (AD) are the impairment of memory related cognitive functions (Hodges, Salmon, & Butters, 1992; Nebes, 1992). Often these symptoms go unreported until those suffering from AD are either pressured by family into getting tested or their level of impairment causes disruptions in their daily lives. Unfortunately, by the time the impairment has affected their daily lives there is little chance of improvement, making early detection of AD much more crucial. To assess memory related cognitive functions clinicians have adopted the use of cognitive tests developed by memory researchers. In return these cognitive tests have given clinicians the opportunity of diagnosing earlier stages of AD, which allow for early interventions that afford patients more control over the progression of AD.

A prominent cognitive test used by medical doctors to measure memory related deficits is a free recall task. The design of the task involves a study trial, where words are sequentially presented to the participant; followed by a test trial where the participant is asked to recall as many of the presented words as they can. Despite its simplicity, the free recall task provides a way to test the strength of a participants episodic memory for familiar words presented on a study trial. A U-shaped serial position curve is often observed from the recall behavior

in a free recall paradigm, where words presented at the beginning and end of the study-list have a higher probability of being recalled than words in the middle of the list (Murdock, 1962). The two peeks of the U-shape are commonly referred to as the primacy and recency effects, respectively.

Generally, two different cognitive processes are assumed to underlie these effects. For the primacy effect, the increased recall probability is assumed to be due to an additional amount of rehearsal time allotted for encoding into a long-term episodic memory (LTM) system (Rundus, 1971). During this additional time, words presented at the beginning of the study list have fewer competitors to rehearse and encode than words presented later in the study list. Additionally a shared characteristic of words that are successfully encoded into LTM is a recall probability that decays slowly over time. As for the recency effect, words toward the end of the study list do not have the extra time available for rehearsal before the test-trial is administered, so the augmentation in their recall probability is thought to be a function of a different system. Namely, words toward the end of the study list may be in a temporary memory system that affords direct access for recall. This system is often referred to as the short-term memory (STM) storage, and words in the STM have recall probabilities that decay rapidly with time. Thus in the STM, words whose proximity is closest to the test trial are more likely to be recalled than items further away from the test trial. Therefore, the recency effect is thought to be a function of STM (Howard & Kahana, 2001). This interpretation is supported by studies where the period between study and test is occupied with a distracting task (Bjork & Whitten, 1974). In this case, words at the end of the study list do not show a recency effect and may result with lower recall probabilities than for words presented earlier in the list.

A widely used test to assess AD related deficits is the Alzheimer's Disease Assessment Scale: Cognitive subscale (ADAS-Cog; Chu et al., 2000; Graham et al., 2004). The ADAS-Cog includes a free recall subtest administered to patients as part of their assessment. The goal of

this paper is to provide a new, model-based assessment method to analyze data in the ADAS-Cog free recall task. For this purpose, the paper develops a cognitively grounded Hidden Markov Model (HMM). The remainder of the paper will be as follows. The next section will review a few operationally defined methods that clinicians have used for analyzing specific latent memory processes in free recall with the progression of AD. This review prompts the need to establish a formal cognitive psychometric model that combines known memory theory to assess the latent processes associated with the free recall task. To do so, the next section will provide specifications and predictions of our model. Next, a method section will provide a description of the research design and data gathered from three groups of participants by the Alzheimers Disease Neuroimaging Initiative (ADNI) using the ADAS-Cog free recall task. In the same section, estimation theory of our model will be presented. Following these sections, preliminary results will provide evidence showing the need for a modification of the model to further facilitate its use in clinical assessment. Finally, there is a discussion of the results and conclusion.

3.2 Clinical Assessment Using Free Recall Data

An important aspect of using the free recall paradigm in populations showing memory deficits is the finding that serial position effects are sensitive in differentiating healthy participants and those suffering from dementia (Egli et al., 2014; Howieson et al., 2011). For example, testing patients with AD related deficits has revealed significant decline in the primacy effect (Capitani et al., 1992; Gibson, 1981). This decline is supported by known LTM deficits associated with the progression of AD and is thought to be due to an impairment of encoding items into LTM. A standard operational method used to measure LTM related processes from the primacy effect is simply to calculate the proportion correct on the first few items in a study list. Researchers are then able to test whether there is a significant

difference between healthy and AD participants using conventional statistics such as the analysis of variance (ANOVA).

While the proportion correct for some items at the beginning of the list has been used as a proxy for LTM strength, other methods have been proposed. A systematic approach using a Selective Reminding Test involves measuring LTM by counting the number of words continually retrieved without further presentation (Buschke, 1973). While this method involves an experimental manipulation different from that of the ADAS-Cog, there have been cognitive models for the Selective Reminding Test (Kraemer et al., 1983; Wenger et al, 2012). Another method of measuring LTM abilities stems from studies of STM on the recency effect. Waugh and Norman (1965) proposed a method that uses performance on the middle words of the list as a proxy of LTM ability, with the assumption that STM processes do not influence the words in the middle serial positions. Regardless of the methodology used, overwhelming evidence for deficits in LTM is reported for patients showing symptomatology of AD (see Carlesimo & Oscar-Berman, 1992, for a review).

The second latent memory process associated with the serial position curve is retrieval from a short-term memory system. Similar to the primacy effect, the recency effect is measured by calculating the proportion correct for a pre-specified number of words at the end of the study list (Tulving & Patterson, 1968). Unlike the primacy effect, clear evidence of recency impairment in AD is not always demonstrated. For example, Martin et al. (1985) and Miller (1971) reported finding a significant reduction in the recency effect for participants with AD etiology. However a study by Spinnler et al. (1988) found normal levels of the recency effect for patients showing signs of AD progression when restricting the analysis of the recency effect to only the last five words. Similarly, Bayley et al. (2000) reported normal recency effects in AD patients when the analysis only included the last two words. On balance, no definitive conclusion can be made about the decline in the primacy effect in AD.

Other, more sensitive, methods have been proposed to measure the latent memory processes

associated with the recency effect. For example, Tulving and Colotla (1970) proposed measuring STM using the performance scores of items with a relatively small distance between the presentation at study and a recall during the test phase. Results using this procedure show comparable STM ability for AD patients and healthy participants for the last 2-3 items in the study list with a significant reduction in STM ability for AD patients on items further away from the test (Carlesimo et al., 1996, Wilson et al., 1983). Methodological differences and severity differences may be the reasons behind the variability in results. However, without a standardized procedure, measurement of these latent memory processes is dependent on the number of words a researcher deems to be part of the primacy or recency effects. By employing formal cognitive models, cognitive psychologists have focused on modeling latent memory structures and processes to improve clinical measures of free recall (e.g. Batchelder et al., 1997; Brainerd et al., 2014). For instance, Batchelder et al. (1997) developed a cognitive model that identified differing cognitive processes underlying the free recall task of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD; Fillenbaum et al., 2008). Batchelder et al. (1997) demonstrated that it was possible with the model to measure differences between AD and cerebrovascular etiologies using the immediate free recall portion of the task. Successful applications of other cognitive models for psychodiagnostic purposes are evident with the many publications of articles in special issues of Psychological Assessment (Neufeld, 1998), Journal of Mathematical Psychology (Neufeld & Townsend, 2010), and chapters in special books on clinical modeling (e.g. Neufeld, 2007).

Although improvement has been made using memory based measurement tools for clinical assessment, often clinical tests used for psychodiagnostic assessments use different task designs that are at variance from those normally used to study memory processes by experimental psychologists. As discussed in Batchelder (1998), the design of assessment tests, like the ADAS-Cog free recall task, is structured so that all participants receive exactly the same test and testing procedure. In contrast, psychological experiments generally control for possible confounding variables known to cause spurious results since episodic memory is

sensitive to experimental design. For example, to avoid item effects, words are chosen to be unrelated to each other and their presentation order on any given study-trial is randomized over participants. On the other hand, the ADAS-Cog free recall subtest is similar to many other clinical tests in that every participant receives the same set of ten words over the same three fixed shuffled order study-trials. The added complexity can be problematic for formal cognitive models attempting to quantify the latent memory processes in the ADAS-Cog free recall test. Any model attempting to analyze such data would have to distinguish between underlying signals from noise created by experimental conditions that do not control for confounding variables.

One solution often used by clinicians is to assume that the added noise occurring from methodology is constant across all participants and thus the true cognitive ability of a person can be approximated by a statistic of their observed responses (e.g. normally the number of correct recalls). In fact, the manual for ADAS-Cog provides scoring rules that create an aggregate summary score for the free recall subtest to diagnose patients showing early signs of AD. With the summary score, researchers can then take advantage of statistical models such as ANOVA to analyze the differences between participant groups. In the results section, we provide a between group repeated measures ANOVA analysis of ADAS-Cog data (see Table 2) on the observed recall behavior to show results using a standard statistical method. Our inclusion of this analysis is designed to point out that summary scores used to quantify behavior on the ADAS-Cog free recall task not only fail to tap into much of the signal in the data, but also do not measure the latent cognitive processes underlying the behavior of those tested on the ADAS-Cog. Instead of analyzing aggregate performance scores, this paper develops and applies a formal modeling based approach that combines known memory theory for a more complete assessment of latent memory processes associated with the free recall task.

3.3 A Hidden Markov Model for Free Recall

The framework for the model in this paper can be traced back to established cognitive models designed for list memory experiments. These memory models stem from the class of models called Hidden Markov Models (HMMs) whose structure involves latent (unobservable) cognitive memory states and observable response sequences. Starting in the 1960s, HMMs became a popular approach to cognitive modeling that led to a number of models that successfully fit data in simple memory paradigms such as paired-associate learning and free recall, (e.g. Greeno & Bjork, 1973; Wickens, 1982). In these memory models, learning is represented as a function of storage and retrieval processes from latent memory encoding states.

In the case of a multi-trial memory task such as the free recall task, a HMM model postulates that on any trial a to-be-remembered item occupies one of a small set of memory states. Associated with each memory state is a retrieval parameter representing the probability of a correct recall for any item occupying that state on a test-trial. The role of the study-trials is to prompt transitions among the memory states through a network of state-to-state transition probabilities specified in terms of the model parameters. Such a model is called a HMM because the observable recall/not-recall response sequence for an item over test-trials does not uniquely identify (hides) the sequence of underlying latent memory states behind the observed response sequence. For example, an error on a test trial could come from any of several memory states. The term Markov comes from a class of stochastic processes where transition probabilities between the states depend only on the current state and not on previous state transitions.

3.4 Basic Model Assumptions

The proposed HMM for the ADAS-Cog recall task postulates three memory states corresponding to different levels of episodic memory storage. The first cognitive state is the Unlearned state (U-State) that represents a state where the participant has not yet encoded a word into episodic memory. The second cognitive state is the Intermediate state (I-State). The I-State is analogous to STM. It is a state where a word is encoded at a shallow level, and the probability of retrieval from that state is expected to decrease rapidly since the occurrence of encoding. The third and final state is the Learned state (L-State). The L-State can be thought of as LTM because it represents a state where an item is fully encoded into episodic memory and it is expected that the recall probability of words in the L-State is subject to slow decay.

It is common practice to display a HMM as a graphical representation of nodes and connections between nodes. The nodes represent model states and the directed connections between nodes represent transition probabilities. The model is represented pictorially in Figure 1. The parameter r is the probability that a word in the U-State is encoded into the L-State on any study-trial. If some encoding occurred but did not result in a transition into the L-State then, with probability $(1 \quad r)a$, the word transitions into the I-State. If no transition from the U-State is made into either of these states then, with probability $(1 \quad r)(1 \quad a)$, the word remains in the U-State. Now if a word is in the I-State at the start of a study trial it has probability of making a transition to the L-State, and with probability it remains in the I-State. Finally once a word is in the L-State, it does not make any further transitions.

The observation recall sequence for the model is the compilation of the recall performance on each item across the four test-trials of the free recall task. These recall events are generated probabilistically as a function of the state an item is in on a test-trial. In Figure 1, the recall probabilities are written inside the nodes for each state. If an item is in the U-State on a

test-trial there is a zero probability of recall, if in the I-State the recall probability is t for immediate test trials, and if in the L-State it is l. The model as currently specified has five parameters, r, a, v, t, and l_1, that represent various transition and recall probabilities.

3.5 Adapting the HMM to the ADAS-Cog Task

To adapt the model in Figure 1 to the ADAS-Cog free recall task, three important additional specifications of the model are needed. First, if an item is in the L-State on any of the three immediate test-trials, the recall probability is l_1; and if an item in the L-State is recalled on a delayed trial (the fourth test-trial), it has a probability l_2 of doing so. Having two recall probabilities for the L-State stems from memory research showing that a memory trace in the LTM decays after a delay (e.g. Burgess & Hitch, 2006). Naturally one would expect that l_2 ¡ l_1 because of memory decay during the delay before the fourth test trial. Unlike the L-State, the I-State is a short-term memory system with rapidly decreasing recall probability, so it is assumed that there is a zero probability of recall on the delayed test trial. Thus the model assumes that only items in the L-State have a chance to be recalled on the delayed test trial.

Second, in addition to the transitions made possible during the study-trial, depicted in Figure 1, there is one other way that a state transition can occur in the model. In particular, if an item is in the I-State on any of the first three study-trials and if, with probability t, it is successfully recalled on the following test-trial, then a transition to the L-State during the test trial is possible with probability b, and with probability (1-b) the item remains in the I-State. This additional transition parameter represents the possibility of learning during the test-trial which is related to the Testing Effect (Goldstein, 2010) and memory research in paired associate learning has shown that learning can occur during a test-trial for both healthy and memory impaired participants (e.g. Bozoki et al., 2006).

The third additional specification of the model concerns how the parameters are tied to the presentation order on a study-trial. When models like the one in Figure 1 are applied to many list memory experiments, a convenient assumption has been to apply the cognitive processes envisioned by the model to each item independently and with identical values of the model parameters. The assumption of identical model parameters regardless of the location of an item in the study list is directly inconsistent with known results provided by the serial position curve. Consequently, our third modification of the model is to adapt the model to the variable study list orders by associating the state-to-state transition parameters and state recall probabilities in Figure 1 with each possible word order position on a study list. Thus the transition and recall probabilities that apply to any word on a particular study- or test-trial depend on the location of that word in the study order for that trial. Since the serial positions for each word change across the study-trials, it follows that any given word will have different state-to-state transition probabilities on each study-trial and different recall probabilities on each test-trial depending on its study list position.

This modification means that for each parameter type in the model, with two exceptions, there is an associated set of ten different parameters, each corresponding to one of the ten study order positions in the 10-word study list. The two exceptions are on the b parameter (learning on a test trial) and l2 (delayed recall from the L-State) because neither parameter is tied to the study list order. However, because the parameter b is a measure of possible learning effects of each item on a test trial, the parameter is indexed by item rather than list order. Consequently there are also ten possible values for the parameter b. The purpose of these generalizations of the model in Figure 1 is to allow the storage probabilities and recall probabilities to reflect the cognitive processes behind the serial position curve discussed earlier. For example, we would expect the r parameter for storage in the L-State in Figure 1 to be higher for a word presented in a position at the beginning of the study list than at the end because of the extra amount of encoding time. Additionally, we would expect the recall probability t from the I-State to be higher for words studied at the end of the list due

to their proximity to the test-phase.

3.6 Methods

Data used in the preparation of this article come from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For current information, see www.adni-info.org.

3.6.1 Participants

ADNIs first longitudinal study (ADNI-1) recruited 744 participants from 50 sites in the United States and Canada. Participants enrolled in ADNI-1 were between 55 and 90 years of age. Normal control participants, (N = 205), had no memory complaints, a Mini-Mental State Exam (MMSE; Folstein, Folstein, & McHugh, 1975) of 24-30, a Clinical Dementia Rating (CDR) of zero, non-depressed, non-MCI, and non-demented. MCI participants, (N = 362), have had a memory complaint by the participant or their partner, MMSE 24-30, objective memory loss based on education adjusted scores on the Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, preserved activities of daily living, and non-demented as determined by the site physician at the time of screening. Mild AD participants, (N = 177), had

MMSE scores between 20-26, CDR of 0.5 or 1.0, and met the criteria established by the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimers Disease and Related Disorders Association (ADRDA) known as the NINCDS-ADRDA criteria for probable AD.

3.6.2 Materials and Procedure

ADAS-Cog was developed in 1983 (Mohs, Rosen, & Davis, 1983) and revised in 1997 (Mohs, et al., 1997) in order for trained personnel to assess cognitive functions affected during the dementia stage of AD using a single, aggregate summary score. The 11 subsections used by ADNI-1 (in no particular order) are: 1) orientation to date and time; 2) constructional praxis; 3) following commands; 4) following multistep instructions; 5) object naming; 6) ideational praxis; 7) spoken language; 8) word finding; 9) word recognition; 10) immediate and delayed free recall; and 11) delayed recognition memory. The data obtained from a 10-word list immediate and delayed free recall tasks is used for the current analysis.

The immediate free recall task has three study-trials where ten words are presented to the participant one at a time on a white index card. During the first study-trial, participants were instructed to read each individual word and to repeat it aloud. After the ten words were studied, the participants were asked to recall the words just presented to them, in any order, within a 2-minute window. The administrator of the ADAS-Cog would then record each word correctly recalled by the participant. This procedure was repeated two more times with the same ten words but with different presentation orders. Each participant received the same instructions and the same material. See Table 1 for the list of ten words and their presentation order on the three study-trials. After the three study-test trials, two intervening tasks were administered followed by the Delayed Recall task. The two intervening tasks tested the participants ability to follow commands and constructional praxis. Neither

Order	Trial 1	Trial 2	Trial 3
1	Butter	Pole	Shore
2	Arm	Letter	Letter
3	Shore	Butter	Arm
4	Letter	Queen	Cabin
5	Queen	Arm	Pole
6	Cabin	Shore	Ticket
7	Pole	Grass	Engine
8	Ticket	Cabin	Grass
9	Grass	Ticket	Butter
10	Engine	Engine	Queen

Table 3.1: Studied Words in the Order Presented to the Participants on Each of the Three Study Trials

intervening task had any common elements with the word recall task. The Delayed Word Recall task tested the participants ability to recall words after intervening tasks. The task required that each participant recall the ten words studied during the three study-trials after a delay of approximately five minutes.

3.6.3 HMM Equations

Discrete trial HMMs are traditionally specified with matrices representing state-to-state transition probabilities, an initial starting state probability vector, and a vector representing state-to-observable response probabilities (e.g. Wickens, 1982). For our current model, two transition operators are required for the proposed design. The first transition matrix, , in equation 1, designates the possible transitions on a study-trial from one of the row states into one of the column states.

The subscript, i, on the parameters refers to the ith study list position on a study-trial, since there are different transition probabilities for each study list order position. Without the study list order subscripts, the transition matrix in equation 1 is just another way of representing the transition probabilities in Figure 1. Note that we have added a column

$$\mathbf{T} = \begin{array}{c|ccc|cc} & \mathbf{L_{n+1}} & \mathbf{I_{n+1}} & \mathbf{U_{n+1}} & \text{Immediate} & \text{Delay} \\ \hline \mathbf{L_n} & 1 & 0 & 0 & l_{1,j} & l_2 \\ \mathbf{I_n} & v_i & (1-v_i) & 0 & t_l & 0 \\ \mathbf{U_n} & r_i & (1-r_i)a_i & (1-r_i)(1-a_i) & 0 & 0 \end{array} \quad (1)$$

vector of recall probabilities, given the row state, to the right of the transition matrix, . The two recall probabilities in the L-State correspond to the immediate test and the delayed test.

The second state-to-state transition matrix, Γ, found in Equation 2, is used to account for the possibility of learning on a test trial.

$$\mathbf{\Gamma} = \begin{array}{c|ccc} & \mathbf{L_{n+1}} & \mathbf{I_{n+1}} & \mathbf{U_{n+1}} \\ \hline \mathbf{L_n} & 1 & 0 & 0 \\ \mathbf{I_n} & t_i b_k & (1-t_i b_k) & 0 \\ \mathbf{U_n} & 0 & 0 & 1 \end{array} \quad (2)$$

Unlike the first transition matrix, T, which covers transitions on any one of the three study-trials, Γ covers possible transitions during any one of the three immediate test-trials. Such test-trial transitions are only possible when an item is in the I-State at the beginning of a test-trial, and with probability $t_i b_k$ the word k at position i transitions into the L-State after the test phase. This probability represents the joint probability of recalling an item located in the I-State and transitioning to the L-State on the test-trial. In the proposed HMM, this is the only process that results in learning during the test trial.

To complete the HMM and obtain the equations needed for statistical inference using the likelihood function of the model we designate the initial start vector of state probabilities before the first study trial as $\Lambda = [0, 0, 1]$, which indicates that every item is assumed to be

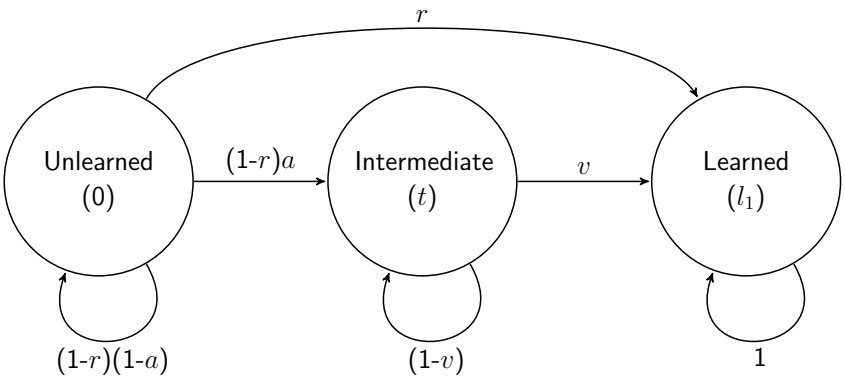

Figure 3.1: Hidden Markov model with the state-to-state transition and state recall over the three latent memory states. State-to-state transition probabilities are written next to the arrows and recall probabilities are written in the circles that represent states.

in the U-State before the first study trial. Now, it is a property of the HMM structure that the probabilities of every sequence of observable responses (recall success or failure) can be obtained by suitable matrix operations on the start vector, transition matrices, and recall vector, e.g. Wickens (1982). There are sixteen possible sequences of observable responses for each of the ten words. For example, it is possible for a participant to fail to correctly recall a word on all four test-trials (0000) or recall that word for all four trials (1111), or anything in between. The observed recall performance sequence for each participant and each word constitutes the basic data that will be used to estimate the parameters of the model.

3.6.4 Hierarchical Bayesian Inference

A central task of psychological assessment focuses on the evaluation of the individual patient's performance or lack thereof. While many cognitive based models are in the service of studying memory related functions at the group level (see Batchelder, 1998, for a review), a resolution towards analyzing individual performances is needed for psychodiagnostic assessments. One standard method to augment a statistical model to handle individual differences is to make it hierarchical. The approach makes the assumption that every participant's parameters are a sample from a hierarchical population distribution with its own parameters

that provide a mean and variance. Given a group of people classified with similar symptomology, estimation of population level latent variables can shed some light on the disorder. For example, given the three groups of participants in the current study, clinicians may be interested in knowing the average latent ability associated with encoding into a long-term storage system for each stage of memory impairment. In addition, the approach of making the model hierarchical allows the analysis for individual participants within a group.

The use of Bayesian based inference for parameter estimation in hierarchical models is an established practice in statistics, e.g. Bernardo and Smith (2000). In the last decade it has become an increasingly popular approach to estimating hierarchical cognitive models, e.g. Lee and Wagenmakers (2013). The advantages over classical likelihood based analysis are both practical and conceptual (Gelman, et al., 2013). From a pragmatic point of view, easily available software such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) and JAGS (Plummer, 2011) allows users to estimate complex cognitive models with relative ease. At the conceptual level, Bayesian statistical inference facilitates the augmentation of formal mathematical models to include hierarchical assumptions to estimate participant parameters. By augmenting the model to include hierarchical assumptions and adopting the Bayesian statistical inference framework, we sidestep many problems posed by classical likelihood methods such as the assumption that the data constitute a large sample of independent and identically distributed observations.

In order to apply hierarchical Bayesian estimation to the HMM, we augment each parameter to have a hierarchical distribution, where individual participant parameters are drawn from a population distribution specific to each parameter. A popular hierarchical population distribution used by statisticians is the Gaussian distribution with mean μ and precision $1/\sigma^2$ hyperparameters. Of course, values sampled from the Gaussian distribution are on the real line, so for our application values drawn from a Gaussian distribution will require a transformation to the probability space of (0,1). A common transformation that takes

values on the real line to values in probability space is the inverse-probit transformation (Gowans, et al., 1989). A graphical model of the hierarchical HMM is presented in Figure 2. Graphical models are helpful to conceptualize the hierarchical structure of the model using nodes and edges corresponding to random variables and their statistical relationship to each other, respectively (Lee & Wagenmakers, 2013). Typically, square nodes indicate discrete variables and circular nodes represent continuous variables. A single border on a node represents stochastic variables and nodes with a double border are deterministic. The shaded nodes represent observed values and replications of portions of the graph structure are enclosed within rectangles.

For the current application, draws from a distribution for the mean and precision of the Gaussian hierarchical distributions are on the real line and on the positive half, respectively. We selected the hyperprior for the mean, , to be normally distributed with the mean and precision set at 0 and 1, respectively. This hyperprior is exactly the distribution of the probit of an uninformative uniform distribution on the probability space (0,1). For the precision hyperparameter, a Gamma distribution is used with scale and shape hyperpriors set at 5 and 5, respectively. The use of uninformative hyperparameter distributions is selected so that before evidence of the data, no parameter value is expected to be more likely than any other value, thus allowing the data to drive the posterior parameter distributions rather than our prior beliefs.

The supplementary material provides the equations for the sixteen response patterns for each of the ten words along with the model likelihood function written in terms of JAG's code. The analysis of the hierarchical model in JAGS used 4 chains of 1,000 samples each with a burn-in of 500 samples. A collective total of 2,000 samples were retained for the current analysis. For a detailed explanation of MCMC sampling, see Gelman et al., (2013). In the results, the reported means over participants of each parameter will be presented in the figures on the natural probability scale rather than on the real line. To obtain the mean of a

particular parameter in (0,1), first an inverse probit is taken of each draw for that parameter from the hierarchical Gaussian, and then the posterior mean and standard deviation of these transformed draws are presented in the figures.

3.7 Results

Before analyzing the recall data with the HMM, it is useful to inspect several aspects of the data. Figure 3 provides the group average recall probabilities for each of the ten words and four test-trials in each of the three groups. The most obvious fact about these plots is that performance decreases across the three study groups. In addition the bar plots for Trial 1 in Figure 3 tend to reveal the expected form of the serial position curve. The remaining bar plots in Figure 3 for the other trials illustrate the effects of the staggered order of words in the study-trials. For all three participant groups, the U-shaped serial position curve is not evident after the first test trial. By changing the presentation order of the words in the study list for the second and third trials as shown in Table 1, words may no longer be governed by the same STM and LTM processes affecting them on the first trial. Consequently, there is no reason to expect the U-shaped serial position curve on those trials.

Although the bar plots for Trial 1 tend to show the expected serial position curve, there are noticeable exceptions, with the largest being in position seven. The word 'Pole' in position seven on the first study-trial has consistently higher recall probabilities than its neighbors for all three-study groups. For some reason, 'Pole' is more memorial than its neighbors in the context of the particular words and word orders that are fixed for all participants in the ADAS-Cog task. Violations of the expected serial position curve such as for item 7 are a likely consequence of the fact that the ADAS-Cog task does not counterbalance the assignment of words to study positions as is done in most experimental studies of free recall. This is an example of one of the difficulties in applying cognitive models to data from clinical

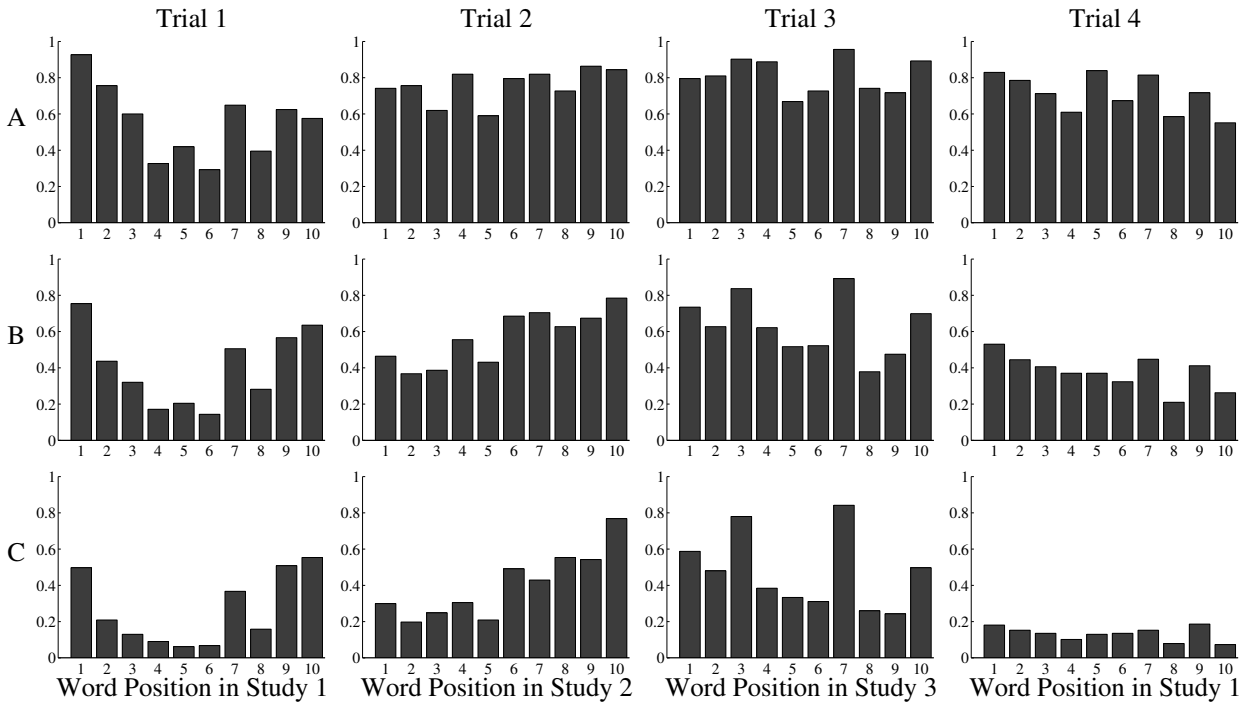

Figure 3.2: The aggregate recall probability for (A) healthy, (B) mild cognitive impairment (MCI), and (C) Alzheimers disease (AD) participants for each word over the four trials. The words are positioned to reflect their assignment during the three study trials, and the fourth trial matches the first trial order.

assessment batteries (see Batchelder, 1998). Then if every participant receives the same set of words in the same order, like in the ADAS-Cog test, specific item effects may become evident such as the word 'Pole' in the ADNI data.

3.7.1 Standard ADAS-Cog Analysis

Next we examine the recall data by employing standard statistical tests on the aspects of the data that are suggested by the ADAS-Cog manual. Table 4 shows the results of a split-plot repeated measures ANOVA (Kirk, 1968, pp. 248-251). The repeated measures factor was the recall trials for the Healthy, MCI and AD groups, and the dependent variable was the number of correctly recalled words. Group and trial main effects and group trial interaction effect significantly influenced word recall performance (p ¡ .01), such that increasing

Source	SS	*df*	MS	*F*	Prob ¿ *F*
Between subjects	12718.29	743	–	–	–
Group	5878.33	2	2939.16	318.41	¡0.01
Subjects w/in Groups (error)	6839.96	741	9.23	–	–
Within subjects	6266.26	2232	–	–	–
Trial	3115.31	3	1038.44	869.13	¡0.01
Group Trial	495.00	6	82.50	69.05	¡0.01
Trial Subjects w/in Groups (error)	2655.95	2223	1.19	–	–
Total	18984.54	2975 6.38	–	–	

Table 3.2: Split-Plot Repeated Measures Analysis of Variance (ANOVA) of the Number of Words Recalled in Each of the Four Trials by Impairment Group

severity results in lower numbers of correctly recalled words; especially on the delayed recall trial. The effect size for Group was calculated using eta-squared, , which by conventional standards is considered large. One difficulty with this type of analysis is with the aggregation of word recall into a single score. Aggregate scores do not offer much insight into the psychological differences underlying the three groups. While the ANOVA model provides useful information, it does not indicate whether normal aging, MCI or AD differentially influences particular cognitive processes that underlie the performance on different words in the aggregate score.

3.7.2 Participant Heterogeneity

To test participant heterogeneity in light of possible item heterogeneity, a nonparametric Monte Carlo permutation test in Smith & Batchelder (2008) was employed. In that article, their test was applied to free recall data obtained from a study similar to the ADAS-Cog free recall task. The test calculates the variance of the participants performance scores on each of the four test-trials. The permutation test obtains a distribution of these variances under permutations of the performance data across participants, and this distribution represents the variability of this statistic under the null hypothesis of participant homogeneity within a study group. The current application tests the null hypothesis that subject variability

for the ADAS-Cog data on each trial is what would be expected from random error. A sample of 100,000 permutations provided a distribution of possible variances of participants performance scores on each trial under the null hypothesis of participant homogeneity. The observed variance for each group across the 4 trials and the 95-percentile distribution of possible variances under the null hypothesis is provided in Table 3. The null hypothesis of participant homogeneity was rejected because the p-values were outside the .05-level (2-tailed) for all four trials for all three-study groups, with the exception of the first trial of the Healthy participant group. The results indicate that it is important to utilize an estimation method for the HMM that handles random effects on the parameters due to participant heterogeneity within a study group. The hierarchical Bayesian inference discussed earlier is ideal for accomplishing this purpose.

3.7.3 Preliminary Results of the HMM

The presence of several defining characteristics from memory theory was discovered in an initial application of the model to the ADAS-Cog data. Two psychological phenomena in free recall described previously were the primacy and recency effects. As mentioned before, memory theory dictates that the primacy effect is reflected by a system responsible for encoding words into long-term episodic memory storage. In the current HMM, encoding into a long-term memory storage system corresponds to the r parameter, which indicates the transition probability into the L-State from the U-State. Figure 3 shows the mean parameter estimates on the probability scale for each serial position averaged over all participants along with a one standard deviation bar from the Bayesian analysis for the Healthy, MCI, and AD participants. Figure 3.A reveals the pattern suggested by memory theory, namely parameter r estimates belonging to the beginning of the list tend to be larger than those towards the end of the list for all three participant groups. Additionally there are large drops in the r parameter with increasing levels of dementia, especially for the early study list positions.

It is important to emphasize that these results, suggesting a primacy effect in parameter r, were in no way forced by the methodology used to estimate the parameters.

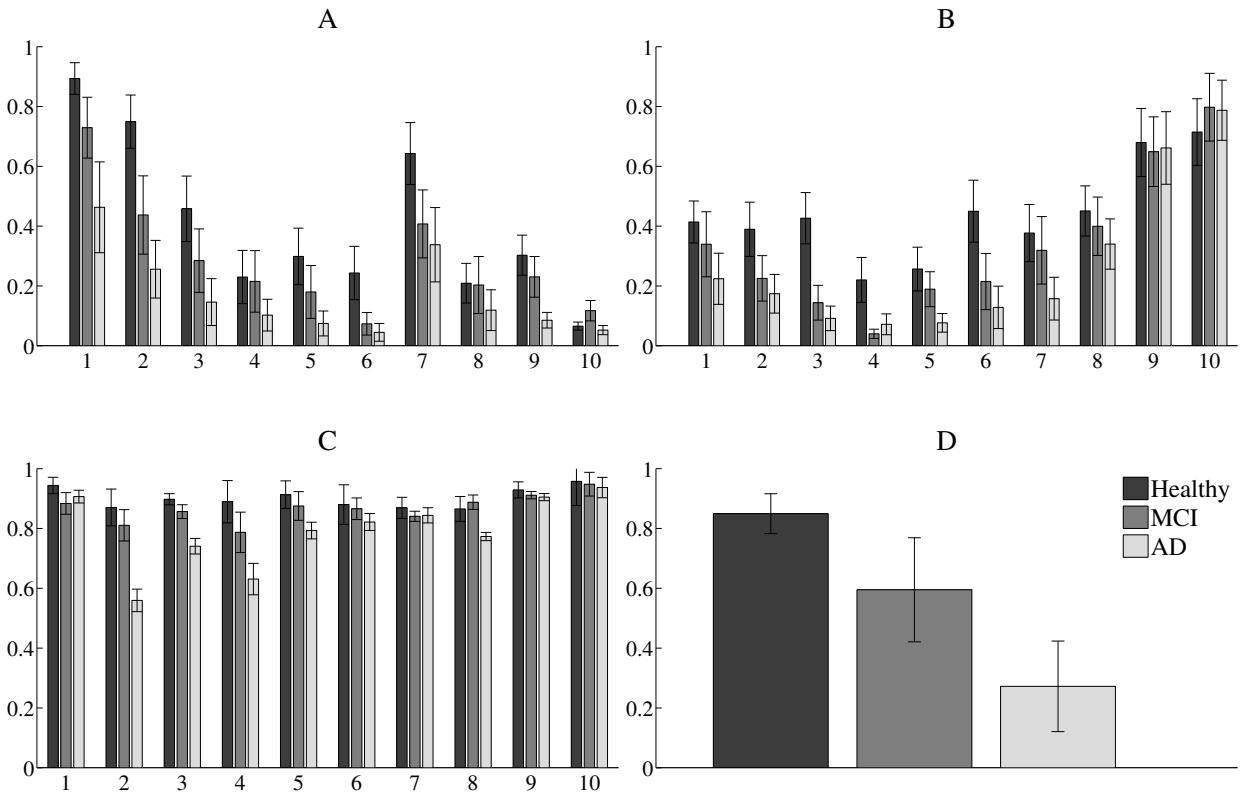

Figure 3.3: (A) The mean r storage parameters (error bar: 1 SD) for the 10 serial positions for healthy, mild cognitive impairment (MCI), and Alzheimers disease (AD) groups, obtained from the posterior distributions. (B) The mean t retrieval parameters (error bars: 1 SD) of the 10 word list positions, for healthy, MCI, and AD groups, obtained from the posterior distributions. (C) The mean l1 retrieval parameters (error bars: 1 SD) of the 10 word list positions for healthy, MCI, and AD groups, obtained from the posterior distributions. (D) The mean l2 retrieval parameter (error bars: 1 SD) for the healthy, MCI, and AD groups.

Figure 3.A does have exceptions to the strictly decreasing pattern one would expect to see for the primacy effect. The most prominent exception is in position seven of parameter r for the three participant groups. As noted earlier there is some inherent unsystematic noise in the data shown in Figure 2 that can be attributed to experimental procedures. Unfortunately, this inherent noise now seems to be carried over into the parameter estimates of the model showing that the words in positions seven have a higher probability of entering LTM than expected by the amount of rehearsal time allotted to those positions. While no model can

Group	Trial 1	Trial 2	Trial 3	Trial 4
Healthy	2.355 (1.7466, 2.5212)	2.500*(1.4610, 2.1081)	2.353 *(1.2059, 1.7452)	3.457 *(1.6136, 2.3
MCI	2.667 *(1.7419, 2.2959)	3.326 *(1.9578, 2.5838)	3.497 *(1.8186, 2.3892)	5.399 *(1.9745, 2.6
AD	2.253 *(1.3100, 1.9349)	2.805 *(1.7137, 2.5432)	3.636 *(1.7036, 2.5447)	2.358 *(0.9260, 1.3

Table 3.3: Permutation test (and 95% Confidence Intervals) for the Three Groups

be assumed to be completely correct, for the purpose of understanding the latent memory processes, an evidenced-based revision of our model will be proposed to control for effects that may arise from the use of testing procedures that do not counterbalance items and their order.

Now the recency effect is thought to be based on a systematic retrieval process from STM. Figure 3.B provides the mean value from the population posterior distribution for the t parameters on the probability scale along with one standard deviation bars for each study list position corresponding to the Healthy, MCI, and AD participants. The parameter t represents a retrieval probability from the I-State, which our model assumes is a short-term storage system with a rapidly decaying memory trace similar to the theoretical STM. This suggests that retrieval from this memory storage should reflect the memory theoretic predictions of the recency effect. Figure 3.B shows that the estimates of t tend to increase with proximity to the test phase for all three-participant groups. Furthermore, there is evidence that the overall recall probability from the I-State decreases with levels of dementia further away from the end of the list. It can be noted that there exists some deviations from an expected recency effect; in particular it appears that the parameters for serial positions at the beginning of the list have slightly higher recall probabilities relative to positions in the middle of the list than might be expected from memory theory. Again, a revision of the model will be introduced to standardize the signal from the unsystematic noise.

The final psychological phenomena discovered in the analysis of the HMM is found in the L-State retrieval parameter. Figure 3.C presents the mean l1 values for each serial position with one standard deviation bars on the probability scale. As a reminder, parameter l1 is the

probability of recalling a word from the L-State on a test-trial presented immediately after a study-trial. Memory theory suggests that when words have been encoded into a long-term episodic memory system, the recall probabilities would not have a rapid decay. Consequently, it is consistent with memory theory to expect that the recall probability l1 would not show a presentation order effect because the time between study and test is relatively short for each study list position. This property is found in the estimates of the recall parameter for the L-State, shown in Figure 3.C. In particular, values of l1 are similar across different word positions for the three groups of participants.

An additional characteristic of the estimates of l1 is the lack of a sharp decline in the MCI and AD groups as found for the other parameters. This is interesting because it shows that not all latent cognitive processes specified in the model have equivalent amounts of deficit across various levels of cognitive impairment. However, the level of retrieval from the L-State after a delay, measured by l2, shows the difference a time delay can have on different levels of deficits. Comparing Figures 4.C and 4.D shows that there is a relatively small decline in retrieval from the L-State for the healthy participants as expected from memory theory, while also revealing a large drop in retrieval ability from the L-State for the AD patients. As for the remaining parameters, no discernable trend was observed across the ten serial positions. Table 4 displays the means of each of these parameters averaged over the10 serial positions along with one standard deviation obtained from the average posterior distributions of each participant parameter. Differences between groups are seen in the parameters, a, and v but not in b.

In summary, the analysis of the ADAS-Cog data with our model has revealed patterns in the parameters r, t, l1, and l2 that are consistent with memory theory, which provides some construct validity for the model and indicates that it may be useful for interpreting the differences in the groups as shown by the simple ANOVA in Table 2. Of course the patterns of these estimates were not in perfect accord with expectations from memory theory, and

Variable	Healthy	MCI	AD
a	0.6967 (.1630)	0.6240 (.2043)	0.4934 (.1790)
v	0.3457 (0.1785)	0.1641 (0.0844)	0.1465 (0.0803)
b	0.3986 (0.1439)	0.3377 (0.1818)	0.3552 (0.2223)

Table 3.4: Average Parameter Values for a, v, and b

we attribute this to a combination of ordinary random variability as well as a result of the experimental design, where all participants had the same list of words in the same set of trial-to-trial orders. Modifications to the model in the next section will focus on strengthening the signal and eliminating noise produced by the experimental procedures.

3.7.4 Evidence-Based Revision of HMM

The goal of the preliminary analysis in the previous subsection was two-fold. The first goal of the analysis was to accentuate the similarities between memory theory and the results obtained by analyzing the data with the model. In this way, evidence for construct validity of interpreting the parameters as tapping latent cognitive process was obtained. The second goal, assuming success of the first, was to discover which latent cognitive processes are affected by increasing levels of impairment. To complete the second goal, it would be beneficial if the cognitive measurements were less affected by the fixed structure of the ADAS-Cog experimental design. To do this the relationship between memory theory and our model parameters is explored and further strengthened by adding constraints on the parameters to match known psychological phenomena. By adjusting our model's parameters to match memory theory assumptions, we create a cognitive psychometric model whose application gives more interpretable measurements of the latent processes that are affected by increasing levels of impairment.

For the current data we will modify three parameters in order to get more interpretable results. Based on the patterns discernable in Figures 3, we add parameter specifications to

r and t by requiring the underlying parameters to satisfy weak order constraints. The order constraints on both r and t are as follows: if j ¡ k then rj rk and tj tk. Imposing these order constraints on r and t does not reduce the number of parameters just their relationships to each other. This approach has been used in another cognitive psychometric models applied to special clinical populations, (e.g. Riefer et al., 2002) . The third modification will be on parameter l1. The modification of the parameter l1 is based on the findings in Figure 3.C that show little difference across the ten l1 parameters. As a consequence we equate the ten l1 parameters within each study group over the study list positions. The remaining parameters that reflect study list positions, a and v, showed no discernible patterns so we imposed no constraints on them, and in addition the parameters b and l2 were not constrained.

The hyperparameter distributions of the unconstrained parameters (ai, vi , bk, l2) in the model are set as before to be drawn from independent Gaussian distributions with mean 0 and precision 1, as is the single l_1 parameter in the constrained model. These draws for each participant are transformed via an inverse probit as before, and then for the figures they are averaged to create values on the probability scale. In the case of the order constrained parameters r and t, one addition to the sampling scheme for the unconstrained HMM is needed to impose the order constrains. Order statistics (David & Nagaraja, 2003) on parameters r and t are applied after the inverse probit transforms. In particular, an inverse probit is applied to each participants set of draws for the ten t parameters, and then they are ordered from smallest to largest. The means in the figures are based on the participants means of these ordered draws. This approach in essence assumes a uniform distribution on all ordered sequences $0 \leq t1 \leq \ldots \leq t1 \leq 1$ making them equiprobable. The same approach generated the distribution of the ten r parameters, except the study position subscripts are reversed. The likelihood function for the modified model has the same functional form as the original model, but its domain is restricted to parameters that satisfy the constraints of the modified model. In other words, it is the restriction on the hierarchical population distribution samples that assures that the posterior distribution of the modified parameters

will satisfy the model constraints. The modified model is analyzed with JAGS and the model code can be found in the supplementary section of this paper. For the following figures, the parameter value reported is obtained from the average posterior distribution means of each participants individual model parameters. One standard deviation bars will be presented to indicate the dispersion of parameter values across participants for each group in the current study.

3.7.5 Results of Modified HMM

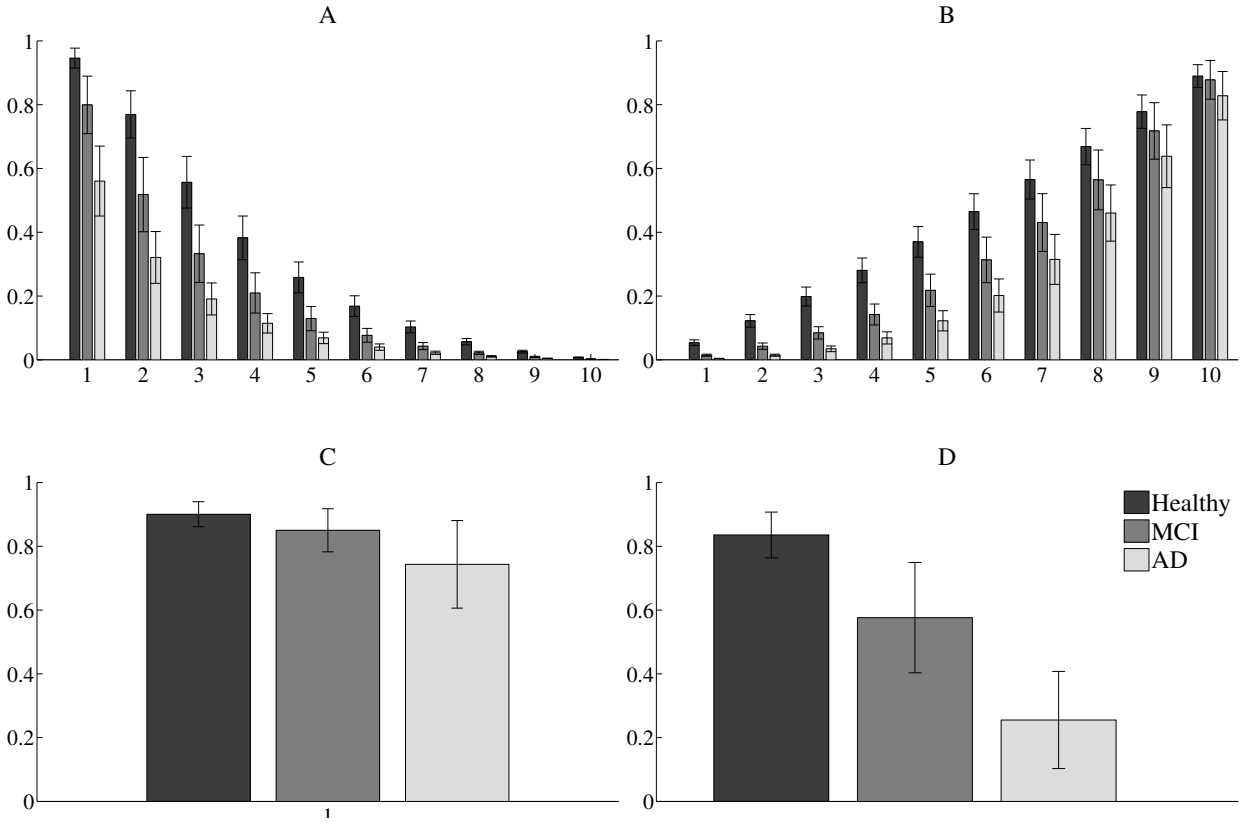

Figure 3.4: (A) The mean r storage parameters (error bars: 1 SD) of the 10 serial positions for healthy, mild cognitive impairment (MCI), and Alzheimers disease (AD) groups, obtained from the posterior distributions of the order constraint model. (B) The mean t retrieval parameters (error bars: 1 SD) of the 10 word list positions for healthy, MCI, and AD groups, obtained from the posterior distributions of the order constraint model. The mean l1 (C) and l2 (D) retrieval parameters (error bars: 1 SD) for healthy, MCI, and AD groups, obtained from the posterior distributions of the order constraint model.

Figure 4 provides the results of applying the modified model to the data. The first result concerns the parameter r. Figure 4.A reveals that words located at the beginning of the study list have a higher probability of being stored into the L-State than words presented at the end of the list, which of course reflects the effect of imposing order constraints. What is of interest is the difference in the r parameters between the three participant groups. Figure 4.A shows that the Healthy group has the highest value of the r parameters followed by the MCI group and then the AD group. It appears that the decline in ability to transition a word into the L-State for the MCI participants is closer to the Healthy group for the beginning two positions and after the third position it drops to similar levels as for the AD participants.

The first retrieval probability of interest is the probability of recalling a word from the I-State, namely the parameter t. The probability of recalling a word from this temporary storage state for each position for the order-constrained model is in Figure 4.B. Again we see the imposed order effect on the parameter t. Figure 4.B shows that words in the last two serial positions are about equally likely to be recalled by a participant in any of the three groups. The groups begin to diverge in their recall ability at position 7. The recollection by AD participants was the first to drop at the 7th position followed by MCI participants at the 6th position. The Healthy participants probabilities of recall from the I-State remained above the MCI and AD group for the preceding positions.

The two retrieval parameters governing the recall probability from the L-State are presented in Figure 4.C and 4.D. The immediate recall from the L-State during the study-test portion of the task, l1, shows that the three groups have similar rates of retrieval in the L-State as was also seen in the unmodified model. The difference between the three groups shows that not all the cognitive processes are affected by the progression of dementia. The largest difference between the three groups on any parameter is shown in Figure 4.D for the parameter l2, which represents the probability of recalling a word from the L-State on the delayed test. The recall proportion after the delay is much lower for the AD group than the MCI group.

Now, when a word is not encoded and stored into the L-State from the U-State, the system allows encoding to occur by three other processes. The first process is conditional on the word not being encoded into L-State, in this case with probability a the word can be encoded into the I-State. No discernible patterns were noticed across the ten positions for any particular group, however, the estimates showed a difference between the groups. Table 5 presents the average value of the ten a parameters for each of the three groups. This shows a decreasing pattern across the three groups in ability of encoding into the I-State. A second method of encoding a word into the L-State is through the parameter v that encodes a word from the I-State into the L-State during a study-trial. Once again, no detectable patterns were observed in each group across the ten serial list positions; however, MCI and AD participants showed a decreased probability compared to the Healthy group. Finally, the b parameter in Table 5 represents the third way a word can be encoded into the L-State from the I-State. The transition is only possible during a correct recall of a word from the I-State on the test-trial with probability b. The result in Table 5 for parameter b shows that learning during the test-trial is not reduced for the MCI and AD groups. As a generalization, it is noteworthy that the results for these parameters in Table 4 match those of the modified model in Table 5.

3.7.6 Discussion of Model Results

Analysis of the ADAS-Cog free recall data with the HMM revealed several interesting explanations behind the significant differences between the three participants groups found in the ANOVA test. The measurement model revealed that, compared to Healthy participants, MCI participants show an impairment in some but not all of the latent memory processes. Furthermore, the latent memory processes found to be impaired in the MCI group, showed a greater degree of impairment for the AD group. The remainder of this section will be devoted at discussing the differences between the MCI group and the AD group.

Variable	Healthy	MCI	AD
a	0.6645 (0.1617)	0.5925 (0.2194)	0.5452 (0.2369)
v	0.3921 (0.1939)	0.1905 (0.1218)	0.1940 (0.1165)
b	0.4581 (0.2176)	0.4535 (0.1885)	0.4419 (0.2842)

Table 3.5: Average Parameter Values for a, v, and b for the Modified Model

By operationally defining latent variables, previous research has shown that encoding into a long-term episodic memory system is impaired with the progression of AD. In our current analysis using a cognitive model, we observe the same decline with the progression of AD. This is important because the model unlike the operational approach of calculating recall proportions for various items combines interacting memory processes that are simultaneously at play during the free recall experiment. A comparison between the three groups reveals a progressive decline in a patients ability to encode information into a long-term storage system. Although the decline is most notable for the AD group, the MCI group does not show an equivalent ability of encoding as the healthy participants. It seems that the encoding process is affected at an early stage of AD such that the system responsible for encoding items into LTM shows a marked decline for words in later serial positions compared to words at the beginning of the study-list. In other words, MCI participants appear to be able to encode information into their episodic LTM provided that few items are competing for encoding.

As for words not encoded into LTM by the MCI group, their conditional transition into the STM as measured by a is nearly has high as the Healthy participants. This finding suggests that a person diagnosed with MCI can still store information into a short-term memory state as well as Healthy participants. However, with a fast decay rate in the STM it seems necessary that the memory trace be encoded into a longer, more permanent system in order to have a greater chance at remembering at a later time. To see if this is indeed the case, we can look at the v parameter in the model, which corresponds to a transition from STM to LTM. It appears that for the MCI group that words encoded into STM are no more likely

to transition into the LTM as in the AD group. This result suggests that the cognitive process used to encode words into the LTM from the STM is hindered during early stages of AD. As outlined by Gauthier et al. (2006), MCI is defined as the prodromal stage of AD where participants are classified by an inability to recall conversations or recent events. This matches the current finding of a large drop in ability to encode words into a more permanent memory structure from a temporary one. Mainly, if the memory trace of a conversation or recent event cannot be encoded into the LTM, the chances of retrieving that conversation or recent event is quite low after some time has lapsed.

The final encoding parameter described by the model that allows participants to encode a word into LTM is the parameter b. The parameter b is an item specific parameter that shows that storage for salient items may occur if they were recalled during the test phase. Similarly, patterns across the first trial for the three groups in Figure 2 showed that certain items were more likely to be recalled across the three groups, such as the word Pole. This suggests that memorable words are as likely to be recalled for healthy participants as for AD patients. The parameter b shows similar values across the three groups, suggesting that encoding during the test phase may be item dependent, and not completely impaired in the progression of AD.

Next we focus our discussion on the recall probabilities of the latent memory states. While overall there is a decline in t with impairment, it is not evident either for MCI or AD participants in t for words at the end of the study list whose proximity to the test phase is closest. One possibility for this result is that even with increasing levels of cognitive impairment, words in STM presented right before recall are still resonating in the STM and compete equally well for recall with other memory traces for all three groups. This interpretation is consistent with the findings of Bayley (2000) and Carlesimo et al. (1996) showing that the last 2-3 items are equally recallable. STM impairment in the progression of AD can be viewed as an inability to retain words further away from the end of the list. In

other words, for MCI and AD participants, memory is subject to a faster rate of forgetting in STM as shown by a large drop in retrieval ability for words further away from the list.

Despite the deficits in recall outlined so far, not every cognitive processes is immediately diminished in MCI and AD participants. The result for the parameter l1, reflecting immediate recall from the L-State, is quite interesting. Basically, using both the original and revised HMM, it was shown that recall probabilities from the LTM on immediate test trials were independent of serial position. In addition, there did not seem to be any noticeable effect of impairment level on l1, and if this result holds up in other applications of the model, it represents a new finding about episodic memory deficits. In particular, even though there is considerable impairment in achieving a long-term episodic memory trace as measured by r and v, there is no deficit in the ability of that trace to support recall when the recall test occurs soon after the encoding.

In contrast, the parameter l2 that measures the delayed recall probability of the L-State, shows a large decline with increasing levels of impairment. This finding shows that after approximately five minutes AD participants ability to recall from LTM diminishes very fast. It seems that as a patients impairment level increases their ability to retain information is no longer aided by what should be a long-term and slow decaying memory state. This interpretation of the L-state comes from the results of the healthy group, showing a small decline in their retrieval ability. While the relatively short time span between an initial measure of LTM retrieval ability and the delayed retrieval measure should not warrant such a decrease, it has been noted that forgetting occurs fastest after short time lag for AD participants (Hart et al., 1987).

	Groups	Trial 1	Trial 2	Trial 3	Trial 4	Total sum
Healthy	M1	0.9610	0.9024	0.8244	0.9220	0.9463
	M2	0.9707	0.9024	0.8293	0.9220	0.9659
MCI	M1	0.9641	0.9392	0.9088	0.9807	0.9448
	M2	0.9696	0.9696	0.9503	0.9807	0.9614
AD	M1	0.9548	0.9887	0.9266	1.000	0.9379
	M2	0.9718	0.9887	0.9887	1.000	0.9774

Table 3.6: Proportion of Bayesian p-Value's within the Corresponding 95% Credible Interval

3.7.7 Assessing Model Adequacy

The new constrained HMM reflects knowledge of the latent cognitive processes based on psychological theory. To test whether the modifications do not limit the ability of the model to fit the observed data, we test the fit of the unconstrained model and the constrained model. To test each hierarchical model with the data, we use a Bayesian p-value (Gelman et al., 2014) on the posterior predictive distributions. First one selects a statistic of the data that is deemed important. Then a distribution of this statistic is generated from various parameter sets obtained from samples during the Markov Chain Monte Carlo runs. Each such sampled parameter set is used to simulate a data set from the model, and from each such data set the value of the chosen statistic is obtained. A distribution made of these samples constitutes the posterior predictive distribution of the statistic, and it can be thought of as the distribution of future data conditioned on the model posterior parameters. Then a p-value for the statistic is obtained by referring the observed value of the statistic to this distribution.

A testable statistic that is often used for free recall is the number of correctly recalled words over the 4 trials. For example, as mentioned the ADAS-Cog manual recommends analyzing the number of correctly recalled words for each participant on each trial. For the current test, the fit of each model will be evaluated on these statistics for each participant. Thus distributions of replicated values predicted by each model for the number of correctly recalled words on each trial and the total number correct across all trials were computed for each

participant. Then the p-value for each participant and trial is the location of the observed data in the distribution of posterior predictive replicated values (Gelman, Meng, & Stern, 1996). Table 6 shows the proportion of participants whose p-values lie within the 95% probability interval of the distribution of replicated values, which is known as the predictive concordance of the model. Support for a model is indicated when this value is near to 95% (Gelfand, 1996). Both models performed fairly well using this test since the total sum values are not far from the desired value of .95, which shows that the modifications did not impair the models ability to fit the data. In the current study, the goal of the comparison is not to deem one model version to be better than other; rather, in any particular application one can decide whether the original model or the order-constrained version is the better way to analyze the data. The modification of the model is motivated by psychological theory and the check of fit to the observed data shows that the changes do not create a worse fit. By focusing on model fit, the Bayesian p-value can be used to measure the discrepancy as a measure of model adequacy (Meng, 1994).

Now that the hierarchical model has shown the ability to account for the observed aggregate recall scores, it is of interest to see if certain individual model parameters can perform well in explaining variations across participants within a study group in the observed scores. Of course there are many model parameters that the model combines to achieve the fits reported in Table 6, so we selected two central parameters as candidates to study, namely l1 and l2. In addition we selected the proportion of correct recalls on the first three test trials (score between 0 and 30) and the proportion correct on the delayed trial (score between 0 and 10). Table 7 presents the Pearson product-moment correlation coefficients of three comparisons.

It is noted in Table 7 that the correlations between l2 and trial 4 performance scores are highest in each group. The model assumes that correct recall is possible on the delayed trial only if the item has reached the L-State, so this is a nice predictive result for the model. Note, performance on a delayed recall test has been shown to be sensitive in differentiating

Group	$\rho(l_1, l_2)$	$\rho(l_1, T_{1-3})$	$\rho(l_2, T_4)$
Healthy	.3970*	.6328*	.8281*
MCI	.2395*	.4811*	.8970*
AD	.3004*	.7245*	.9530*

Table 3.7: Pearson Product-Moment Correlation Coefficients Between Long-Term Memory (LTM) Retrieval parameters L_1 and L_2 and Performance Scores on the First Three Trials

AD and Healthy participants (Welsh et al., 1992), so l2 may help explain the differences between the groups. Now, the correlations between l1 and the first three test trials are lower. Since performance on the first three test trials can come about both from the I-State and the L-State, one would not expect that l1 would be able to explain as much variance in the first three trials as does l2 for the delayed trial. The correlation between l1 and l2 is consistent with the model assumption that l1 has decayed during the delayed test. Consequently participants with a larger l1 are more likely to have a larger value of l2 after the delay regardless of their group assignment.

3.8 Conclusion

Early detection of AD is quite important to clinicians and families of those with the disease. An advantage that clinicians have is their use of cognitive tests to classify the likelihood a person is impaired. These cognitive tests are similar to the experimental procedures used to develop cognitive models to measure latent memory processes. For this reason it makes sense to attempt to apply cognitive models rather than simple statistical analyses to data obtained from the cognitive tests used by clinicians. By combining a formal cognitive model with established psychological theory, we are able to measure a few latent cognitive processes associated with learning and memory from the behavioral measures collected by clinicians using the ADAS-Cog free recall task. Doing so allows a more complete picture of which cognitive processes are affected by the progression of dementia.

The class of Hidden Markov models is adopted to accomplish our purposes. Our HMM demonstrates that with a simple two memory storage state system the primacy and recency effects are a byproduct of underlying latent cognitive processes extrapolated from behavioral measures gathered using a variable order study list. Support for the two memory systems has been provided by both memory theory and neuropsychological studies showing different memory disorders associated with deficits in performance for early and later parts of the study-list (Basso et al., 1982; Baddeley & Warrington, 1970).

The result of the analysis demonstrates that by using a mathematical model, we are able to circumvent potential problems caused by having to estimate where the primacy effect ends and where the recency effect begins. An immediate consequence of this can be seen with the measurement of the recency effect. The result of our model supports the findings that, although AD participants show a decline in recall from STM, their problem arises from an inability to recall earlier items but not the last items in the study list. In effect, reducing the STM capacity of storage, possibly causing a diminished ability found in the recency effect. Another finding, after application of the model, was seen in the retrieval process in the long-term storage state, showing similarity to the hypothesized LTM. Further, the retrieval characteristics of this state, as measured by the Healthy group, indicated a level of forgetting consistent with a slow decaying process. The rapid forgetting measured for the AD group indicates that AD patients ability to hold on to memory rapidly declines after an initial storage (Hart et al., 1987). The comparison between Healthy and AD was possible because of similar initial immediate recall abilities from LTM, as measured by the parameter l1. The effects of this decline is possibly the reason behind the low performance on delayed trials and may be the reason why the delayed recall task is sometimes used as a proxy for measuring LTM (Welsh et al., 1992).

One motivation for the current application of the model to the three groups was to establish the utility of standardizing measures of cognitive function in the progression of AD, any of

which might be impaired with the progression of AD (Nebes, 1992). Other applications of our model can prove useful, for example, by quantifying the latent variables, clinicians may be able to use the model for assessing changes, if any, in psychological processes with certain drug interventions. Other applications of the model can include the study of memory disorders, such as vascular dementia; to understand what latent processes are affected by their disorder. By broadening the scope of the models application to other memory disorders, a comparison across memory disorders can then be attainable. Doing so may reveal additional benefits such as assessing which psychological processes are helped by certain prophylactics.

In addition by making the model hierarchical it becomes possible to use it to detect participants that might be misclassified by physicians since the Bayesian hierarchical inference of the model returns posterior distributions of each parameter for each participant. Nevertheless the estimation theory for the model may have to be augmented to allow it to better classify individual participants. One idea for future work would be to construct an empirical Bayesian prior based on a large sample of participants who enter a clinic and are tested. Such a prior would be highly informed unlike the priors that were used to analyze the HMM. Batchelder (1998) describes how a data bank of this sort could be constructed based on a cognitive model and used to classify individuals.

3.9 References

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long-and short-term memory. Journal of verbal learning and verbal behavior, 9(2), 176-189.

Basso, A., Spinnler, H., Vallar, G., & Zanobio, M. (1982). Left hemisphere damage and selective impairment of auditory verbal short-term memory. A case study. Neuropsychologia, 20(3), 263-274.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. Psychological Assessment, 10(4), 331.

Batchelder, W.H., Chosak-Reiter, J., Shankle, W.R., & Dick, M.B. (1997). A Multinomial Modeling Analysis of Memory Deficits in Alzheimer's Disease and Vascular Dementia. Journal of Gerentology, 52, 206-215.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. Psychonomic Bulletin & Review, 6(1), 57-86.

Batchelder, W.H., & Riefer, D.M. Using Multinomial Processing Tree Models to Measure Cognitive Deficits in Clinical Populations, In R. Neufeld (Ed.). Advances in Clinical Cognitive Science: Formal Modeling of processes and Symptoms. Washington,D.C. American Psychological Association Books, 2007, pp. 19-50.

Bayley, P. J., Salmon, D. P., Bondi, M. W., Bui, B. K., Olichney, J., Delis, D. C., ... & Thal, L. J. (2000). Comparison of the serial position effect in very mild Alzheimer's disease, mild Alzheimer's disease, and amnesia associated with electroconvulsive therapy. Journal of the International Neuropsychological Society, 6(03), 290-298.

Bernardo, J. M., & Smith, A. F. (2009). Bayesian theory (Vol. 405). John Wiley & Sons.

Bjork, R. A., & Whitten, W.B., (1974). Recency-Sensitive Retrieval Processes in Long-Term Free Recall. Cognitive Psychology, 6: 173–189.

Bozoki A, Grossman M, & Smith E.E. (2006). Can patients with Alzheimer's disease learn a category implicitly? Neuropsychologia, 44, 816–827.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. Psychological Review, 113, 201–233.

Brown, G. D., Della Sala, S., Foster, J. K., & Vousden, J. I. (2007). Amnesia, rehearsal, and temporal distinctiveness models of recall. Psychonomic Bulletin & Review, 14(2), 256-260.

Burgess, N., & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. Journal of Memory and Language, 55, 627– 652.

Buschke, H. (1973). Selective reminding for analysis of memory and learning. Journal of Verbal Learning and Verbal Behavior, 12(5), 543-550.

Capitani, E., Della Sala, S., Logie, R. H., & Spinnler, H. (1992). Recency, primacy, and memory: Reappraising and standardising the serial position curve. Cortex, 28(3), 315-342.

Carlesimo, G. A. (1996). Recency effect in Alzheimer's disease: a reappraisal. The Quarterly Journal of Experimental Psychology: Section A, 49(2), 315-325.

Carlesimo, G. A., & Oscar-Berman, M. (1992). Memory deficits in Alzheimer's patients: a comprehensive review. Neuropsychology review, 3(2), 119-169.

Celone, K. A., Calhoun, V. D., Dickerson, B. C., Atri, A., Chua, E. F., Miller, S. L., ... & Sperling, R. A. (2006). Alterations in memory networks in mild cognitive impairment and Alzheimer's disease: an independent component analysis. The Journal of neuroscience, 26(40), 10222-10231.

Chu, L. W., Chiu, K. C., Hui, S. L., Yu, G. K., Tsui, W. J., & Lee, P. W. (2000). The reliability and validity of the Alzheimer's Disease Assessment Scale Cognitive Subscale (ADAS-Cog) among the elderly Chinese in Hong Kong. Annals of the Academy of Medicine, Singapore, 29(4), 474-485.

David, H. A. & Nagaraja, H. N. (2003). Order Statistics. Wiley Series in Probability and Statistics. doi:10.1002/0471722162.

Egli, S. C., Beck, I. R., Berres, M., Foldi, N. S., Monsch, A. U., & Sollberger, M. (2014). Serial position effects are sensitive predictors of conversion from MCI to Alzheimer's disease dementia. Alzheimer's & Dementia.

Fillenbaum, G. G., van Belle, G., Morris, J. C., Mohs, R. C., Mirra, S. S., Davis, P. C., ... & Heyman, A. (2008). Consortium to Establish a Registry for Alzheimer's Disease (CERAD): The first twenty years. Alzheimer's & Dementia, 4(2), 96-109.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state : A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12, 189-198.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., ... & Winblad, B. (2006). Mild cognitive impairment. The Lancet, 367(9518), 1262-1270.

Gelfand A (1996). "Model Determination Using Sampling Based Methods." In W˜Gilks, S˜Richardson, D˜Spiegelhalter (eds.), Markov Chain Monte Carlo in Practice, pp. 145– 161. Chapman & Hall, Boca Raton, FL.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model

fitness via realized discrepancies. Statistica sinica, 6(4), 733-760.

Gibson, A. J. (1981). A further analysis of memory loss in dementia and depression in the elderly. British Journal of Clinical Psychology, 20(3), 179-185.

Goldstein. B. E., (2010). Cognitive Psychology: Connecting Mind, Research and Everyday Experience. Cengage Learning. p. 231

Gowans, E. M. S., Fraser, C. G., & Hyltoft Petersen, P. (1989). A guide to the use of probit transformation of Gaussian distributions. Biochimica Clinica, 13, 327-336.

Graham, D.P., Cully, J. A., Snow, A. L., Massman, P, Doody, R. (2004). The Alzheimer's Disease Assessment Scale-Cognitive subscale: normative data for older adult controls. Alzheimer Disease and Associated Disorders, 18(4), 236-40.

Greeno, J. G. & Bjork, R. A., (1973). Mathematical Learning Theory and the New Mental Forestry. Annual Review of Psychology, 24, 81-116.

Hart, R. P., Kwentus, J. A., Taylor, J. R., & Harkins, S. W. (1987). Rate of forgetting in dementia and depression. Journal of Consulting and Clinical Psychology, 55(1), 101.

Hodges, J. R., Salmon, D. P., & Butters, N. (1992). Semantic memory impairment in Alzheimer's disease: Failure of access or degraded knowledge? Neuropsychologia, 30(4), 301-314.

Howard, M. W., & Kahana, M. J. (2001). A Distributed Representation of Temporal Context. Journal of Mathematical Psychology. doi:10.1006/jmps.2001.1388.

Howieson, D. B., Mattek, N., Seeyle, A. M., Dodge, H. H., Wasserman, D., Zitzelberger, T., & Jeffrey, K. (2011). Serial position effects in mild cognitive impairment. Journal of clinical and experimental neuropsychology, 33(3), 292-299.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. Psychometrika, 59(1), 21-47.

Kirk, R. E. (1968). Experimental Design: Procedures for the Behavioral Sciences. California: Wadsworth Publishing Company, Inc.

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. Journal of Mathematical Psychology, 48(4), 215-229.

Kraemer, H. C., Peabody, C. A., Tinklenberg, J. R., & Yesavage, J. A. (1983). Mathematical and empirical development of a test of memory for clinical and research use. Psychological-Bulletin , 94 (2), 367.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. Perspectives on Psychological Science, 6(3), 299-312.

Lee, M. D., & Wagenmakers, E. J. (2013). Bayesian cognitive modeling: A practical course. Cambridge University Press.

Martin, A., Brouwers, P., Cox, C., & Fedio, P. (1985). On the nature of the verbal memory deficit in Alzheimer's disease. Brain and Language, 25(2), 323-341.

Meng, X. L. (1994). Posterior predictive p-values. The Annals of Statistics, 1142-1160.

Merkle, E.C., Smithson, M., & Verkuilen, J. (2011). Hierarchical models of simple mechanisms underlying confidence in decision makings. Journal of Mathematical Psychology, 55, 57–67.

Miller, E. (1971). On the nature of the memory disorder in presenile dementia. Neuropsychologia, 9(1), 75-81.

Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., ... & Thai, L. J. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. Alzheimer Disease & Associated Disorders, 11, 13-21.

Mohs, R. C., Rosen, W. G., & Davis, K. L. (1983). The Alzheimer's disease assessment scale: an instrument for assessing treatment efficacy. Psychopharmacology bulletin, 19(3), 448-50.

Moore, D. S., & McCabe, G. P. (2006). Introduction to the practice of statistics (5th ed.). New York: Freeman.

Morris, R. G., & Baddeley, A. D. (1988). Primary and working memory functioning in Alzheimer-type dementia. Journal of Clinical and Experimental Neuropsychology, 10(2), 279-296.

Murdock, B. B., Jr. (1962). The Serial Position Effect of Free Recall, Journal of Experimental Psychology, 64, 482-488.

Nebes, R. D. (1992). Cognitive dysfunction in Alzheimer's disease. In F. I. M. Craik & T. A. Salthouse (Eds.), The handbook of aging and cognition (pp. 373-446). Hillsdale, NJ: Erlbaum.

Neufeld, R. (Ed.). (1988). Process Models in Psychological Assessment [Special issue]. Psychological Assessment, 10(4).

Neufeld, R., & Townsend, J. (Ed.). (2010). Contributions of Mathematical Psychology to Clinical Sciences and Assessment. Journal of Mathematical Psychology, 54(1).

Palmeri, T. J, & Flanery, M. A. (1999). Learning about categories in the absence of training. Psychological Science, 10(6), 526–530.

Petersen, R., Thomas, R., Grundman, M., Bennett, D., Doody, R., Ferris, S., Galasko, D., Jin, S., Kaye, J., Levey, A., Pfeiffer, E., Sano, M., van Dyck, C., & Thal L. (2005). Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment. New England Journal of Medicine, 352, 2379-88.

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March (pp. 20-22).

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. Psychological Review, 95(3), 318.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. Psychological Assessment, 14(2), 184.

Rogers, S. L., & Friedhoff, L. T. (1996). The efficacy and safety of donepezil in patients with Alzheimer's disease: results of a US multicentre, randomized, double-blind, placebo-controlled trial. Dementia and Geriatric Cognitive Disorders, 7(6), 293-303.

Rundus, D. (1971). Analysis of rehearsal procedures in free recall. Journal of Experimental Psychology, 89(1), 63-77.

Spinnler, H., Sala, S. D., Bandera, R., & Baddeley, A. (1988). Dementia, ageing, and the structure of human memory. Cognitive Neuropsychology, 5(2), 193-211.

Pepin, E. P., & Eslinger, P. J. (1989). Verbal memory decline in Alzheimer's disease A multiple-processes deficit. Neurology, 39(11), 1477-1477.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. Psychonomic Bulletin & Review, 15(4), 713-731. Chicago.

Tulving, E., & Colotla, V. A. (1970). Free recall of trilingual lists. Cognitive Psychology, 1(1), 86-98.

Tulving, E., & Patterson, R. D. (1968). Functional units and retrieval processes in free recall. Journal of Experimental Psychology, 77(2), 239.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. Psychological review, 72(2), 89.

Welsh, K. A., Butters, N., Hughes, J. A., Mohs, R. C, & Heyman, A. (1992). Detection and staging of dementia in Alzheimer's disease. Use of the neuropsychological measures developed for the consortium to establish a registry for Alzheimer's disease. Archives of Neurology, 49, 448-452.

Wenger, M. J., Negash, S., Petersen, R. C., & Petersen, L. (2010). Modeling and estimating recall processing capacity: Sensitivity and diagnostic utility in application to mild cognitive impairment. Journal of mathematical psychology , 54 (1),73-89.

Wickens TD (1982). Models for Behavior: Stochastic Processes in Psychology. W. H. Freeman and Company, San Francisco.

Wilson, R. S., Bacon, L. D., Fox, J. H., & Kaszniak, A. W. (1983). Primary memory and secondary memory in dementia of the Alzheimer type. Journal of Clinical and Experimental Neuropsychology, 5(4), 337-344

Chapter 4

Retention as a Function of Retrieval in long-term Memory

4.1 Introduction

Among the most important consequences of learning is the formation of lasting memories. Efforts to increase retention levels against the normal course of forgetting often focus on promoting an active involvement, whether by means of rehearsal or testing. While these two interventions are different in many regards, they share the common goal of developing a lasting representation in a long-term memory (LTM) storage system. Unfortunately the preponderance of memory impairments in disorders such as dementia and Alzheimers disease is typically characterized by an abnormal course of forgetting. To mitigate memory impairments caused by dementia, clinicians have relied on a wide range of treatments that include the use of prophylactics to cognitive training procedures.

Before such treatments can be applied, clinical studies are conducted for early detection of memory loss. A prominent psychological test designed to study human episodic memory and

commonly used in conjunction with clinical batteries that assess memory related deficits is the multi-trial item free recall task. The task involves a study phase where participants study a list of words and a test phase where participants are asked to recall the studied words aloud in any order. After several of these study-test phases, recall is again measured after a long delay. While successful recall of the studied items relies on a memory system capable of storing and retrieving information, a precise description of the structure of human memory is unknown. However, certain characteristics have been observed using the free recall paradigm, and in return these reveal some functionality of the latent dynamical system.

Furthermore, there is evidence that cognitive decline in dementia-related illnesses is systematically manifested through progressively declining performance scores on tasks designed to study episodic memory. For example, a problem often associated with memory deficits in Alzheimers disease (AD) is that a residual memory trace is often weaker when compared to healthy controls on tests that were administered after a prolonged period of time from the last encoding phase. The loss of retrievable information from LTM is generally characterized by an initial accelerated forgetting rate, tapering off after a longer period of time (Hart et al, 1987). Results from an application of a hidden Markov model (HMM) to data collected from AD patients have corroborated previous findings that residual memory trace decay may be occurring at a faster rate compared to healthy controls after a five minute delay trial (Alexander et al. 2015).

The purpose of this paper is to further understand retention of memory as a function of overt recall behavior. The target of this study is to find potential treatments capable of improving retention for those suffering from dementia. Using data from a large corpus collected by clinicians for a comprehensive multi-trial free recall task, we focus our efforts on assessing the cognitive processes associated with learning and memory with a hidden Markov model (Alexander et al.). Modifications to the model are needed to resolve dynamic changes that arise from a multi-trial cognitive based test not predicted by linear learning models. Using

deterministic trend-stationarity techniques ensures that the stochastic process underlying the changes in response outputs across trials is identifiable.

The remainder of the paper is organized as follows: First, we introduce the learning model and its properties. Second, the clinical test and data is presented. Following the result section, we discuss our findings in the conclusion section.

4.2 Learning Model

Consider a 3 state HMM for a multi-trial free recall task where only a portion of the observation (test) trials come after a study session. The states space, $\mathbb{S}$, of this discrete HMM model is the set of episodic memory states, $\mathbb{S} = \{U, I, L\}$ (defined below) with distinct retrieval characteristics. Let M be the number of items in a study list, N_s the number of trails that immediately follow a study session, N_D the number of trials that do not have an immediately preceding study session and N be the total number of test trials. A test item, k, where $k = \{1, \ldots, M\}$ is assumed to be in exactly one of the three states as a result of the study.

The first of the three states is an unlearned state (U-State) characterized by a failure to encode retrievable information resulting in an inability to correctly recall test items on the test trial. The next transient state represents a temporary storage state referred to as the intermediate state (I-State). Following a study trial, information stored in the I-State is retrievable with an increasing probability rate favoring items towards the end of the study list. The third and final state is the learned state (L-State) and represents a state where an item is fully encoded into episodic memory. This absorbing state retains information for the remainder of the experiment so it is natural to assume that successful recall on a delayed test is achieved for items stored in the L-state.

A central feature of HMMs is the dependence between states, such that on any study trial, an item follows a Markov process. This process is expressed as the transition operator, where k is the index by test item, as follows:

$$\mathbf{T} = \begin{array}{c} \\ L_n \\ I_n \\ U_n \end{array} \begin{array}{c} \begin{array}{ccc} L_{n+1} & I_{n+1} & U_{n+1} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ v_k & (1-v_k) & 0 \\ r_k & (1-r_k)a_k & (1-r_k)(1-a_k) \end{bmatrix} \end{array}.$$

The index applies a corresponding matrix operator to each individual item in a study list with the assumption that the memory states are states of all items.

The ij^{th} term of the row stochastic lower Hessenberg matrix with rank 3 corresponds to the probability of an item transitioning from a current state i on study trial n to state j on the next study trial, $n+1$, where $S_n = i, j \in \{U, I, L\}$. Beginning from the lower left, the probability of transitioning from U-State to L-State from trial n to $n+1$ is r_k. A failure to transition into L-State from U-State during a study trial but where information is successfully encoded into the I-State is $(1-r_k)a_k$. If no transition from the U-State is made into either of these states then, with probability $(1-r_k)(1-a_k)$, the word remains in the U-State. From the I-State an item can transition into the L-State with probability v_k if there remains a residual trace of the item during the previous study trial, or with probability $(1-v_k)$ the item remains in the I-State. Finally, once a word is in the absorbing L-State, no further transitions are made.

The state-response mapping is given by the response operator, with l_k giving the probability of recall from the L state, and t_k giving the probability of recall from the I state, as follows:

$$\mathbf{R} = \begin{bmatrix} l_k \\ t_k \\ 0 \end{bmatrix}$$

An important distinction between the memory states in the model is the retention interval governing the successful retrieval of a test item after encoding. Aside from the U-State, the response characteristics for each state are probabilistic and subject to temporal changes. The I-State is representative of a short-term memory storage process, thus the successful retrieval of information is inversely related to the interval from encoding to retrieval. The weak storage of the I-State suggests very little resistance to forgetting so the parameter t is indexed by k. Note, index k is appropriate when items do not change positions in a study list across multiple study trials; however, if the items are shuffled then the index represents list position (see Alexander et al, 2015).

Now the L-State, being analogous to long-term memory, does not suffer from the same limitations as the I-State. Rather, once an item is encoded into the L-State, retrieval is the same for each item in the list regardless of its position in the list. Although the retention interval is longer, it is not limitless. We may show this by indexing the retrieval probability with k_d where $k_d = 0$ for test trials following a study trial and $k_d = \{1, \cdots, N_d\}$ for N_d delayed test trials. Retrieval from the U-State and I-State during a delay is set at zero. Finally, to complete the description of the model the start vector is $\pi = [0\ 0\ 1]$. The assignment that at the start of an experiment all items begin in the U-State is made to reflect the fact that before the task, no item is more likely to be part of the study list.

Thus far, the memory model for a multi-trial free recall task is a reparameterization of a two-stage model previously applied to clinical data (Shankle, Batchelder, 1997). The reparameterization decouples the conditional relationship between a general transition out of the U-State and into the L-State. This is done to emphasize a difference in learning as a process of leaving the U-State versus learning as a directed operation into different memory states. Therefore, the transition into L-State from U-State does not depend on the probability of a transition into the I-State.

4.2.1 Multiple Operators

There are two additional aspects of the model that are not revealed by **T**. Evidence by Bjork has shown that when participants are tested, learning is not limited to the study sessions alone. Instead, information is not only retrieved on a test trial but is subsequently enhanced for future recall. Thus, in a repeated measures experiment, success on subsequent trials can be due to an inadvertent enhancement as a result of successfully recalling a word on the previous trial. Termed as the testing effect, the process has a different effect on words that are successfully retrieved from those that fail be to be recalled. To include the testing effect, a separate transition operator is applied that depends on a successful retrieval from I-State. Formally, if $k \in I$-State at n_s and Pr(x_{n_s} =1 | I-State) = t, then Pr($S_{n_s+1} = L$-State | $S_{n_s} = I$-State) = b and Pr($S_{n_s+1} = I$-State | $S_{n_s} = I$-State) = $1 - b$. Remember that the subscript n_s indicates trials in which the test trial occurs immediately after a study trial.

In addition to transitioning into L-State from the I-State during the test trial, the transient state has a susceptibility of failing to encode residual information into an absorbing state. Failure to sufficiently condition residual information into an absorbing state is indicative of a forgetting process. A transition back into the U-State during the study phase is counter intuitive mainly because the learning phase provides an opportunity for an item to transition through memory states. To resolve this, we include a transition back into the U-State after a failure to correctly recall an item from the I-State. More formally, if $k \in I$-State at n_s and Pr(x_{n_s} =0 | I-State) = $1 - t$, then Pr($S_{n_s+1} = U$-State | $S_{n_s} = I$-State) = $1 - \theta$ and Pr($S_{n_s+1} = I$-State | $S_{n_s} = I$-State) = θ.

Both additions represent a transformation too complex for the current model space. For reasons of mathematical convenience the I-State and L-State are split to reflect a single response outcome from each state bifurcation. Altogether the model is comprised of 5 states with 2 states corresponding to successful recall labeled with an asterisk. Using the expanded

matrix notation, the additional transition operator conditioned on the subject response is:

$$\mathbf{\Gamma} = \begin{array}{c} \\ L_n^* \\ L_n \\ I_n^* \\ I_n \\ U_n \end{array} \begin{array}{c} \begin{array}{ccccc} L_{n+1}^* & L_{n+1} & I_{n+1}^* & I_{n+1} & U_{n+1} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ b_i & 0 & (1-b_i) & 0 & 0 \\ 0 & 0 & 0 & \theta_i & 1-\theta_i \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} .$$

The change embeds the 3 dimensional model into a 5 dimensional space without changing the transition process, resulting in a matrix operator that includes the response mapping into the Markov chain. The change creates an unambiguous representation of the complete stochastic process as the single matrix:

$$\tilde{\mathbf{T}} = \begin{array}{c} \\ L^* \\ L \\ I^* \\ I \\ U \end{array} \begin{array}{c} \begin{array}{ccccc} L^* & L & I^* & I & U \end{array} \\ \begin{bmatrix} l_i^* & (1-l_i^*) & 0 & 0 & 0 \\ l_i & (1-l_i) & 0 & 0 & 0 \\ v_i l_i + (1-v_i) t_i b_i & v_i(1-l_i) & (1-v_i) t_i (1-b_i) & (1-v_i)(1-t_i)\theta_i & (1-v_i)(1-t_i)(1-\theta_i) \\ v_i l_i + (1-v) t_i b_i & v_i(1-l_i) & (1-v_i) t_i (1-b_i) & (1-v_i)(1-t_i)\theta_i & (1-v_i)(1-t_i)(1-\theta_i) \\ r_i l_i + (1-r_i) a_i t_i b_i & r_i(1-l_i) & (1-r_i) a_i t_i (1-b_i) & (1-r_i) a_i (1-t_i)\theta_i & (1-r_i)(1-a_i) + (1-r_i) a_i (1-t_i v_i)(1-\theta_i) \end{bmatrix} \end{array} .$$

The rank of the expanded matrix $\tilde{\mathbf{T}}$ does not change, and since the transformation constitutes only a change in basis one can go from $\tilde{\mathbf{T}}$ to $\mathbf{T}$ (see Appendix). One desirable consequence of splitting the states in this fashion is that the otherwise difficult to obtain conditional probability of correctly recalling a word from L-State given a previous successful retrieval, i.e. $Pr(X_i(n+1) = 1 | X_i(n) = 1,\text{L})$, becomes trivial. Thus the probability of retention as a function of previous successful recall from the L-State is given by the piecewise function:

$$Pr(X_i(n+1)=1|X_i(n)=1,\mathrm{L}) = Pr(\mathrm{L}^*_{n+1}|\mathrm{L}^*_n) = \begin{cases} l^* & \text{if } n \leq N_s \\ l^*_k & \text{else} \end{cases}$$

Note that successful retrieval probabilities on delay trials are indexed by the corresponding delay trial. Once again, only items encoded into a strong episodic memory state have a resistance to forgetting. By differentiating retrieval probabilities based on past recall, we can monitor sustainability of memory as a function of past response behavior.

4.3 AVLT DATA

To accomplish the proposed project, I will use a new set of data collected from a unique version of the free recall task used in the Rey Auditory Verbal Learning Test (AVLT). Along with the new data, I will augment the cognitive model developed for a previous clinical free recall task to handle the new data obtained using the AVLT free recall task.

4.3.1 Methods

The data were collected by the Mayo Clinic Aging Study, and contained scores from 178 normal and 131 AD subjects. The AVLT free recall task is composed of five study-test trials and two delay tests on a 15 word list. During the first study-trial, participants were instructed to read each individual word and to repeat it aloud. After the words were studied, the participants were asked to recall the words previously presented to them, in any order, within a 2-minute window. The administrator would then record each word correctly recalled by the participant. This procedure was repeated four more times with the same 15 words

and with the same presentation order for each of the five study trials.

4.3.2 Procedures

Each participant received the same instructions and the same material. After the five study-test trials, an intervening task was administered followed by the first delayed recall task. The first delayed task tested the participant's ability to recall the words after a five minute intervening task having no common elements with the word recall task. Following this 6th test trial, participants completed other unrelated intervening tasks delaying the final recall test by one hour. In total, participant scores to five immediate test trials and two delayed test trials were collected for each of the fifteen items.

The AVLT data set differs from the previous design used to construct our HMM in three important ways. The first difference is that patients are asked to memorize 15 items rather than 10. The increased list size is generally more difficult, so the second modification to the data was to involve 2 more study-test trials, totaling 5 study-test trials. The third modification is the inclusion of another test after a one hour delay. These changes are significant to the augmentation of the new model (e.g., the recall probabilities are negatively influenced by list length). The data for the model are observation sequences of correct and incorrect recall for each of the 15 words. Each word has 2^T possible response sequences where T is the total number of trials. Thus the number of observation sequences for the new set of data is 128 for each item in the study list.

Model Equations

Consider the observed data vector x_n of a learning task where the vector denotes objects either remembered or not $\{1, 0\}$ and the subscript n indicates discrete time in a sequence,

$t = 1, \cdots, N$. The data consists of $O = \{x_1, \ldots, x_N\}$ and sequential dependencies are modeled using a hidden Markov model where the $x_n's$ are stochastic functions of a hidden Markov chain, Q, where q_n is a discrete random variable taking one of E possible values.

The stochastic system for the HMM is a probabilistic process where transition probabilities dictate the next state of the system from step n to step $n + 1$ depending only on the current state of the system. Such a model is called "hidden" because the observable recall/not-recall response sequence for an item over test trials does not uniquely identify (hides) the sequence of underlying latent memory states behind the response sequence. For example an error on a test trial could come from any memory state.

Given an observation sequence (O) of correct or incorrect recall on a fixed number of test trials N, we would like to calculate the probability of such a sequence given the model. Note with a large N it is not feasible to obtain the expressions in closed form; however, with 7 trials the 128 category probabilities will be delineated[1].

First, we enumerate the state sequences,

$$Q = q_1, q_2, \ldots, q_N$$

where q_n is one of the three states i.e. U, I, or L on trial n. The sequence is determined by psychological theory; e.g., once an item is in L-State, it is not possible for the item to return to U-State, and further after the fifth study-test trial there are no further changes in state.

Next, we obtain the probability of an observation sequence given a state sequence[2] is:

[1]Although it is feasible, it is also extensive, for example when $N = 4$, as in Alexander et al., 2015, there are 16 observation sequences that require 136 lines of code for a collective 6 pages of equations.

[2]In order to study the maintenance of memory in the L-State as a function of previous retrieval, $P(O_n|q_n, \mathcal{M})$ must be modified to reflect a dependency between observations on the subsequent test from the L-state with whether or not the word was correctly recalled on the previous test trial while in the L-State,

$$P(O|Q,\mathcal{M}) = \prod_{n=1}^{N} P(O_n|q_n,\mathcal{M})$$

and the probability of the state sequence Q is:

$$P(Q|\mathcal{M}) = \pi_{q_1} \prod_{n=2}^{N} P(q_n|q_{n-1})$$

where π_{q_1} is determined by the start vector and $P(q_n|q_{n-1})$ is a transition probability from state q_{n-1} to q_n. Finally, the probability of an observation sequence is given by summing over all possible state sequences of the joint probability:

$$P(O|\mathcal{M}) = \sum_{\text{all Q}} P(O|Q,\mathcal{M})P(Q|\mathcal{M}) \tag{4.1}$$

The result obtained from equation 1 provides the probability of a single observation sequence out of the 128 mutually exclusive and exhaustive observation sequences. By continuing this process, the remaining probabilities are found, and are quite useful for parameter estimation as we will see in the estimation theory section.

Model Predictions

A central property of learning models is their ability to derive predictions about the relationship between events across trials. Statistics describing the dependency between events

i.e., $P(O_n|q_n,\mathcal{M},O_{n-1})$.

fall under the class of sequential response probabilities which often provide the strongest test of where a model breaks down. Let E_n be the event of an error on trial n, and E_{n+1} be an error on trial $n+1$, then:

For the current model and most learning models $Pr(E_{n+1}|E_n)$ decreases monotonically as n increases.

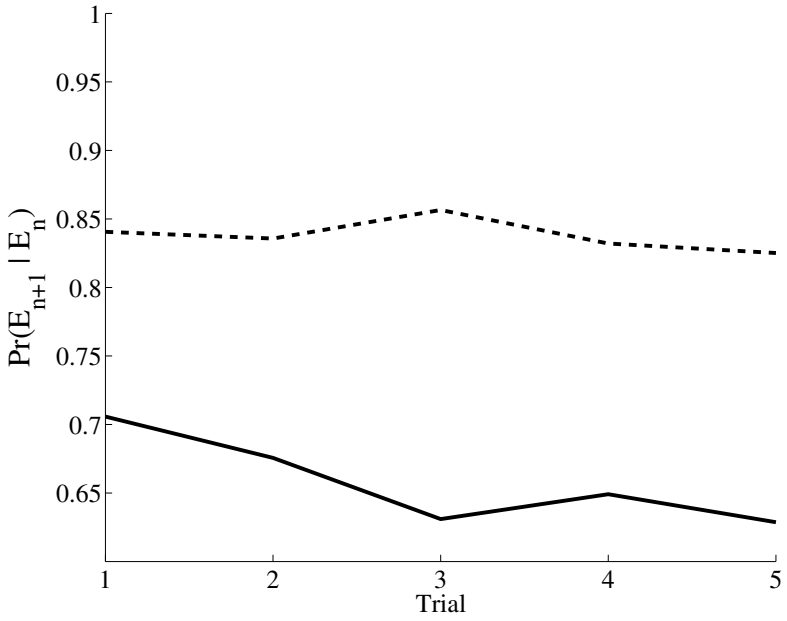

Figure 4.1: Conditional probability of an error on trial $n+1$, given an error on trial n. Top line represents data for AD participants, while bottom line represents data for healthy participants.

Figure 1 shows the conditional error probabilities for the two participant groups on the first 5 study-test trials. The data shows a violation of monotonicity as predicted by learning models. Perturbations of this sort are particularly important to the study of memory using HMM type models since it suggests a nonstationarity problem not easily resolved without the addition of many more states with corresponding theoretical assumptions. Presupposing that the addition of more states can be theoretically defended, methodological problems such as estimating the increased number of independent model parameters would require an unfeasible increase in data. However, since the addition of an unknown number of memory

states is unfounded in theory paired with the methodological issues, another approach at modeling the changes in time is adopted.

While effects of insufficient pre-training mapped by Markov models have been previously observed in research of paired-associate learning (see Kintsch 1963), no suitable solution has been proposed. The problem is further complicated by trying to understand whether the dynamic changes over error rates are a result of changing encoding rates or the result of varying retrieval rates. In other words, the changes may occur because either the retrieval system requires training to acclimate to the more difficult task of retrieving fifteen words, or the process of learning to learn where participants are learning to perform on a novel type of test. The difference is analogous to anchoring either the response operator over time and allowing the transition operator to vary over time, or vice-versa.

One solution that avoids the previously stated problem is to segment the recall events into J subsets of the overall response trials. While this solution may be intuitive, an unequivocal partition of the observation sequences is not possible without increasing the number of parameters exponentially by K^J where K is the number of model parameters. The apparent difficulty of the former solution is tantamount to complications of over-parameterizations which lead to identifiablility problems that result in the misclassification of the relative strengths of cognitive abilities as measured by the model parameters (Bamber & Van Santen, 1985).

Another solution that does not require that we segment trials is to assume the existence of a time-homogenous Markov chain in the strictest sense, i.e., having a stationary transition operator and to systematically appropriate the influence of the operator at every time step

with a deterministic function on time. This is expressed as:

$$\tilde{T}_n = \tilde{T} + \Delta(n) \tag{4.2}$$

where Δ is a matrix of equal size to $\tilde{T}$.

Each transition parameter of the model is assumed to have a separate deterministic function on time.

$$\theta_{i,n} = \Phi(\delta_n + \hat{\theta}_i) \tag{4.3}$$

where δ, $-\infty < \delta < \infty$ and Φ is the cumulative Gaussian function. The probability distribution δ is normal with mean 0 and variance 1. This solution increases the number of parameters by a multiplicative constant.

The application of these additional parameters does not necessarily indicate a fundamental part of the data generating process, but rather is used as an alternative to a more complicated process and unknown function. Furthermore, we do not assume a stochastic trend; instead, a deterministic trend is used since the parameter's variance is not assumed to change across time. If variance is permitted to change across time, the estimates do not converge to a stationary distribution. To test the viability of the additional parameters, the Ruben-Gelman criteria of convergence is obtained for each parameter [CITATION]. The diagnostic criteria often referred to as the rhat statistic evaluates the convergence of the sampling distribution by comparing multiple chains with different starting points for each parameter.

4.3.3 Estimation Theory

There are various methods traditionally used to fit an HMM to data, such as the expectation-maximization algorithm and maximized log-likelihood methods. While both methods are suitable for the current model, the approach we take here is to evaluate the likelihood function using Bayesian inference techniques. The use of Bayesian estimation procedures greatly simplifies the difficulties of obtaining parameter values compared to maximization procedures. Software such as WinBugs (Lunn, Thomas, Best, & Spiegelhalter, 2000) and JAGS (Plummer, 2011) facilitate the matter even further by offering intuitive modeling options.

The likelihood function for the current model takes the form of a multinomial probability distribution. It is easy to see that the probabilities for the 128 mutually exclusive and exhaustive observation sequences obtained by equation 1 can be denoted as the probabilities in a multinomial probability distribution and the observations themselves as the observed outcomes of an experiment. Thus the count for each category probability is obtained from performance scores of each subject for each item. The likelihood function without the multiplicative constant is then:

$$\mathbf{L}(\mathbf{O}, \mathbf{M}) = \prod_{\alpha}^{Ss} \prod_{\beta}^{OS} \prod_{k}^{M} P(O_{\lambda,k}|M)^{O_{\lambda,k}} \tag{4.4}$$

where $\alpha = [1, \ldots, Ss]$ are the subjects, $\beta = [1, \ldots, OS]$ are the 128 observation sequences and $k = [1, \ldots, M]$ are the items in a list.

Along with the likelihood function, specification of the parameter prior distributions allows us to complete the model for Bayesian inference. In Alexander et al. order statistics were applied to a subset of the parameters and their distributions of the model to reflect knowl-

edge about psychological theory. Although the constraints added to the model were at the parameter level, Knapp and Batchelder have shown that this is statistically equivalent to a reparameterization of the model. For the sake of consistency we left the constraints on the parameters used in Alexander et al., The transition probabilities do not depend on item position.

The prior transition probabilities are set to be Gaussian with mean 0 and variance 1. The response probabilities, the prior distributions are set to be uniform between the closed interval [0,1]. This was accomplished with the Beta distribution with shape parameters set at 1,1.

$$\theta \sim dnorm(0, 1) \tag{4.5}$$

$$\hat{\theta} \sim dnorm(0, 1) \tag{4.6}$$

$$\omega \sim dbeta(1, 1) \tag{4.7}$$

The analysis of the model in JAGS consisted of 4 chains of 500 samples each with a burn-in of 500 samples with a total of 2,000 samples. The means over participants of each parameter will be presented on the natural probability scale rather than on the real line. To obtain the mean of a particular parameter in (0,1), first an inverse-probit is taken of each draw for that parameter from the hierarchical Gaussian, and then the posterior mean and standard deviation of these transformed draws are obtained.

4.4 Results

Prior to evaluating the model, a split-plot repeated measures analysis of variance (ANOVA) comparing average performance scores between the two groups was conducted. The repeated measures factors for the analysis of variance are as follows:

Source	SS	*df*	MS	*F*	Prob. $>F$
Trials	75128.8	6	12521.5	19.05	$<.01$
Group	678533.2	1	678533.2	61.59	$<.01$
Interaction	27145.6	6	4524.3	6.88	$<.01$
Subjects (matching)	308464.4	28	11016.6	16.76	$<.01$
Error	110430.4	168	657.3		
Total	1199702.4	209			

The statistical test comparing performance scores across the groups is a good measure of group differences but does not distinguish successful retrieval from short term memory and long term memory.

We found no evidence that the state response parameters changed across study trials, so we only report time dependent parameters for the transition probabilities.

	Healthy						Alzheimer					
Item	a	b	r	t	v	y	a	b	r	t	v	y
1	0.3509	0.1836	0.9855	0.2585	0.058	0.8688	0.1466	0.0071	0.9446	0.156	0.0013	0.5631
2	0.3888	0.5097	0.8316	0.2971	0.1775	0.6976	0.2013	0.0691	0.672	0.2082	0.0081	0.2651
3	0.3347	0.6342	0.5756	0.3338	0.6563	0.2686	0.1785	0.1215	0.3668	0.2542	0.0295	0.0266
4	0.1519	0.1814	0.4983	0.358	0.2981	0.9012	0.0777	0.0137	0.3318	0.3006	0.0543	0.5574
5	0.8896	0.8678	0.4539	0.3782	0.5831	0.497	0.6579	0.6932	0.2854	0.3331	0.2411	0.0432
6	0.2567	0.3253	0.4159	0.3937	0.7344	0.8974	0.1425	0.0225	0.2503	0.3471	0.3927	0.5865
7	0.1739	0.0949	0.3472	0.406	0.4297	0.8258	0.0954	0.0069	0.196	0.359	0.1885	0.4961
8	0.2312	0.2585	0.3104	0.4183	0.7235	0.5477	0.1342	0.023	0.1711	0.3693	0.4482	0.2494
9	0.1774	0.3099	0.2887	0.4323	0.3946	0.7069	0.0987	0.0412	0.1578	0.3806	0.1078	0.3571
10	0.5758	0.8313	0.2706	0.4726	0.6984	0.5198	0.3955	0.628	0.1426	0.4115	0.2788	0.1589
11	0.2875	0.28	0.2516	0.5001	0.4202	0.7517	0.1758	0.0444	0.1313	0.4397	0.1449	0.4211
12	0.7391	0.454	0.2121	0.6803	0.2236	0.0486	0.5986	0.2811	0.1008	0.6388	0.0446	0.0017
13	0.2645	0.0334	0.1625	0.687	0.2808	0.3415	0.163	0.0011	0.07	0.6463	0.0976	0.1183
14	0.746	0.3821	0.0859	0.8144	0.1433	0.0667	0.6179	0.2222	0.0153	0.7811	0.0168	0.0029
15	0.8137	0.496	0.0308	0.9195	0.1214	0.1061	0.6909	0.325	0.0013	0.8919	0.0106	0.0041

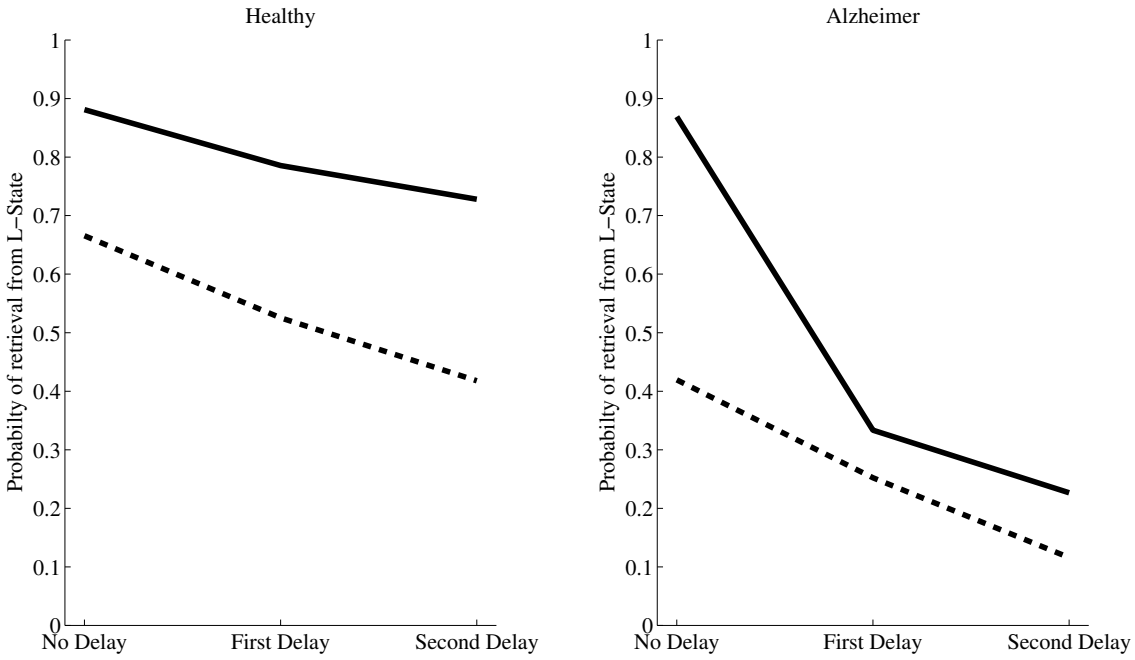

Figure 4.2: Retrieval probabilities conditioned on past recall behavior. Solid lines indicate the probability of retrieving an item from L-State given successful retrieval on previous trial. Dotted lines indicate the probability of retrieving an item from L-State given unsuccessful retrieval on previous trial.

4.5 Conclusion

The prevalence of memory related impairments in dementia patients has given rise to efforts aimed at mitigating the effects of memory decline. With the help of tasks such as those found in clinical batteries we can shed light on the properties often promoted as aiding the retention of information. The active learning task administered to AD patients and healthy controls reveals the benefits of intentional encoding.

While the initial application of the model in Alexander et al. tackled the problem posed by the staggered presentation order used in the ADAS-Cog free recall task the current analysis of the model focused on the generalizing the application to different versions of the same task. In the current application of the model, a non-stationarity effect was discovered and corrected by making the applied parameters functions of time.

Retrieving information from short-term memory acts as a misguided deterrent to continue to strengthen memory. The performance scores show that items situated at the end of the list are less likely to be remembered after a delay. The model shows that those items are recalled with higher probability from short-term memory and are more likely to be returned to the U-State.

Both the design of the task and the circumstances for taking the task invite a person to take the task of learning the material seriously. Thus the responses recorded can shed light on the effects of learning and retention much more than testing college students.

The results show that the successful retrieval of information is not impervious to decay rates. Furthermore, the rapid decay often exhibited by AD participants is unaffected by the success of their encoding.

It is interesting to note that the probability of retrieval from L-State on no delay trials is high and decreases very rapidly, suggesting that machinery used to retrieve information is

uninhibited but the retention of information is damaged.

Future implementation of this model for the data would benefit from a larger sample size.

4.6 Reference

Alexander, G. E., Satalich, T. A., Shankle, W. R., & Batchelder, W. H. (2015). A Cognitive Psychometric Model for the Psychodiagnostic Assessment of Memory Related Deficits. *Psychological Assessment*.

Batchelder, W.H., Chosak-Reiter, J., Shankle, W.R., & Dick, M.B. (1997). A Multinomial Modeling Analysis of Memory Deficits in Alzheimer's Disease and Vascular Dementia. *Journal of Gerontology*, 52, 206-215.

Brainerd, C. J., Reyna, V. F., Gomes, C. F. A., Kenney, A. E., Gross, C. J., Taub, E. S., & Spreng, R. N. (2014). Dual-retrieval models and neurocognitive impairment. *Journal of experimental psychology: learning, memory, and cognition, 40*(1), 41.

Hart, R. P., Kwentus, J. A., Taylor, J. R., & Harkins, S. W. (1987). Rate of forgetting in dementia and depression. *Journal of Consulting and Clinical Psychology, 55*(1), 101.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Chapman & Hall/CRC.

Greeno, J. G. & Bjork, R. A., (1973). Mathematical Learning Theory and the New "Mental Forestry. *Annual Review of Psychology, 24*, 81-116.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing, 10*(4),

325-337.

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *In Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (DSC 2003). March (pp. 20-22).

Nebes, R. D. (1992). *Cognitive dysfunction in Alzheimer's disease.* In F. I. M. Craik & T. A. Salthouse (Eds.), The handbook of aging and cognition (pp. 373-446). Hillsdale, NJ: Erlbaum.

Waugh, N. C., & Smith, J. K. (1962). A stochastic model for free recall *Psychometrika, 27*(2), 141-154.

Welsh, K. A., Butters, N., Hughes, J. A., Mohs, R. C, & Heyman, A. (1992). *Detection and staging of dementia in Alzheimer's disease.* Use of the neuropsychological measures developed for the consortium to establish a registry for Alzheimer's disease. *Archives of Neurology, 49*, 448-452.

Chapter 5

Knowledge Gradient Consensus

"The one thing that does not abide by majority rule is a persons conscience." - Atticus, To kill a mockingbird.

The purpose of the paper is to utilize methodological and mathematical axioms from an established formal modeling approach to data fusion called Cultural Consensus Theory (CCT) in order to extract a more veridical representation of cultural knowledge. CCT was created in the mid-1980s with the combined efforts of scientists in the fields of Anthropology and Mathematical Psychology. The theory relies on the notion that specific cultural knowledge can be studied by asking relevant test items to informants who share that knowledge. Thus far CCT models have assumed that each informant has a unique but measurable competency to correctly answer questions regarding their cultural knowledge. A consequence of this is that if one informant has a higher competency than another about some aspect of their culture, then they are assumed to exceed the other in competency for all aspects of the culture. A more realistic assumption is that there is a division of labor, where each informant has high levels of knowledge in some areas of their shared culture, but not in all areas. The novel extension will create new specifications of CCT models that allow the estimation of

heterogeneous clusters of items and their relationship to corresponding clusters of expert informants. By identifying informant sets corresponding to areas of cultural knowledge, we may be able to better understand the taxonomy of knowledge in a particular culture as well as to reduce uncertainty in the estimation of overall cultural knowledge.

5.1 Introduction

Since Galton's famous prediction of an ox's weight using central tendency measures of individual opinions, scientists have been interested in finding unequivocal procedures to help elucidate the transient nature of truth. While Galton demonstrated it is possible to better approximate the truth using an average response over any individual response, information pooling techniques, such as the majority rule, often employ overly simplified assumptions. Usually, to provide an informed prediction, a researcher (such as Galton) weighs each individual's response equally. In other words, every informant's response is equally likely to be correct regardless of individual differences in expertise. The distribution of knowledge in this work is therefore in a sense uniformly distributed across all informants. However, there are times when a more veridical approximation to the ground truth is critical, for example when deciding which medical intervention to apply in a given situation.

In response to this demand, Cultural Consensus Theory (CCT) was developed in the mid-1980s through the combined efforts of scientists in the fields of Anthropology and Mathematical Psychology (Batchelder & Romney, 1989). CCT is a data fusion technique aimed at constructing an answer key to questions with no known answers, with the help of naturally occurring differences that arise amongst informants. The theory exemplifies the notion that shared knowledge can be objectively measured through the unverified responses of individuals in a culture. Thus, with the additional theoretical sophistication of CCT, researchers are able to correctly infer objective truths from individual responses even when there exists

no clear plurality, as in the case when each response alternative has an equal proportion of votes.

In recent years, an increasing number of papers using CCT have emerged in many fields such as anthropology, sociology, and psychology (e.g. Anders, et al. 2014; Oravecz et al. 2015). So far, each mathematical model developed for CCT has assumed that each informant contains a unique but quantifiable competency to all knowledge within their respective culture, known as item monotonocity (Batchelder & Romney 1989). A consequence of this is that if an informant is considered to know more than another about some aspect of their culture, they are then assumed to know more about all aspects of the culture. A more realistic assumption is that there is a division of labour, where each informant may have high levels of knowledge in some areas, but not all. Clearly, the need for such a distinction can be avoided by generating questions specific to a particular domain. However, as Bradlow, Wainer, & Wang, (1999) pointed out, this introduces a strong relationship between test items which leads to a violation of conditional independence assumed by the model. To include the new assumptions and avoid violations of conditional independent we present a new methodological framework which we call Knowledge Gradient Consensus Model (KGCM).

We begin by reviewing the first CCT model, known as the General Condercet Model (GCM), and explore properties generated from its axioms. Following this, we propose a new modified version of the model that relies on hierarchical assumptions in order to reduce the number of free parameters. Furthermore, the requirement for prior knowledge of item membership to a subset is relaxed and will be explored using a latent mixture model after establishing the utility of such an augmentation. Therefore the new model proposed here assumes there exists quantifiable differences between informants with knowledge about a subset of items from those with knowledge about some other partition of the complete item set.

5.2 Standard GCM

A standard mathematical model for dichotomous data under the aegis of CCT is the GCM. GCM builds upon the assumption that unverified dichotomous responses can be modeled using methodologies in signal detection theory (Anders & Batchelder, 2012). While the usual application of signal detection theory is reserved for cases in which an experimenter contains information about an item and focuses on latent variables underlying the observed behavior, GCM relaxes the requirement of prior item-specific information. In doing so, it becomes apparent that the GCM is a modification of the signal detection theory framework; where instead of assuming knowledge of an item's membership to a signal or noise distribution, it generates an item's membership from the theoretical variables associated with each individual in a culture (Batchelder & Romney, 1988).

Formally, let the response profile matrix be $\mathbf{X} = (X_{i,k})_{N\times M}$ where N is the number of informants and M is the number of items. The GCM is fully delineated through the following axioms:

Axiom 1. (*Cultural Truth*). There exists a single answer key shared by all informants, Z.

Axiom 2. (*Conditional Independence*). The response Profile matrix satisfies conditional independence given by: $Pr[\mathbf{X} = (x_{i,k})_{N\times M}|\mathbf{Z}, \mathbf{H}, \mathbf{F}]$

$$= \prod_{i=1}^{N}\prod_{k=1}^{M} Pr(X_{i,k} = x_{i,k}|Z_k, H_i, F_i) \tag{5.1}$$

for all possible realizations $(x_{i,k})$ of the response profile matrix.

Axiom 3. (*Marginal Probabilities*). The marginal probabilities (2) are given by $Pr(X_{i,k} = x_{i,k}|Z_k, H_i, F_i)$

Axiom 4. (*Double High Threshold*). Hit and false alarm parameters for each informant are

reparameterized by

$$\forall i, H_i = D_i + (1 - D_i)g_i, \tag{5.2}$$

$$F_i = (1 - D_i)g_i. \tag{5.3}$$

The four axioms above formalize the GCM and set testable assumptions needed for the application of the model. From an early start, GCM's success has relied on linear combinations of informant specific competency and bias parameters. In fact, an early method for estimating the parameter values of the model came from minimizing the residuals using methods such as minimum residuals (Batchelder & Romney, 1989) which is a method for finding a numerical solution of a system of linear equations by approximating a solution with minimal residuals. Further studies on factor analysis procedures proposed later as a method for discovering multiple cultures was developed (Batchelder & Anders, 2012).

However, these previous interpretation of the factor-analytic approach to the response profile matrix decomposition has been governed by the belief that the correlation between informants is indicative of cultural membership. While this belief has aided the development of culturally specific answer keys, alternative interpretations of the latent factors underlying the correlation between individuals may help us elucidate a more robust conceptual framework in which to more fully explore additional enhancements to the model. For example, we may reinterpret the correlation between individuals to represent a clustering of responses based on varying levels of ability at the item-specific level.

First, this may represent a more natural psychological way of thinking about the relationship

between informants for an item, where higher levels of agreement between informants may indicate a property of their individual competency rather than attributing all differences in knowledge to a broader cultural membership. Second, this alternative conceptualization may lead to insights concerning additional model enhancements which improve overall model accuracy. When individuals are assumed to have varying abilities for responses on an item, we can see that the abilities in use for a correct response represents knowledge which is declarative in nature, i.e. the individual is able to directly indicate this knowledge in responding to the question. However, this leads to the possibility that an individual may also have some implicit level of knowledge regarding a subject, which is not inherently declarative and therefore may not be evident in the correct response scenario. This implicit knowledge may influence the individual's guessing ability for a given item. The first point will be important when choosing the number of informant clusters and the second point will be explored below in the next section.

5.2.1 Asymmetric Bias Effect

Now, an important contribution of CCT is not just that its mathematical models are able to reconstruct a cultural truth, since majority rule is capable of the same feat, rather the theory's focus on constructing a truth value for an item from informant specific latent measures derived from their response profile. Thus by procuring information from each individual informant in a culture, CCT is able to quantify the degree of agreement across commonalities using their responses to a set of items (Batchelder, 1989). Furthermore, since CCT concerns itself with the prediction of a culturally true answer key from unscored responses it is natural to explore possible predictions made by the theory.

One otherwise unexplored prediction of the model is the posterior probability that the truth value of item k is "yes" given an informant said "yes," i.e. $P(Z = 1|X = 1)$. The conditional

probability $Pr(Z = 1|X = 1)$ reveals the influence of a particular informant on the answer key. Bayes theorem is used to derive the posterior probability of an items truth value conditioned on the informants response.

$$Pr(Z = z|X = x) = \frac{Pr(X = x|Z = z)Pr(Z = z)}{Pr(X = x)}$$

If π is regarded as the prior probability of an item answered "yes," then by substituting terms into the Bayes formula we derive the posterior probability that the truth value of item k is "true" given an informant response "yes."

$$Pr(Z = 1|X = 1) = \frac{\pi p_{1i}}{\pi p_{1i} + (1 - \pi)p_{0i}} \tag{5.4}$$

Now, substituting the terms in eq. 3 with those from eq. 1:2 we get:

$$Pr(Z = 1|X = 1) = \frac{\pi(D_i + (1 - D_i)g_i)}{\pi(D_i + (1 - D_i)g_i) + (1 - \pi)(1 - D_i)g_i} \tag{5.5}$$

Figure 5.1 shows how each parameter in the model differentially acts upon the probability of a latent truth value. More importantly, the plot shows an asymmetric bias effect (ABE) on the conditional probability in favor of an informant with $g = .1$ over an unbiased informant, i.e. $g = .5$, regardless of ability. ABE is a systematic shift in parametric influence to the construction of a latent answer key and it characterizes a redistribution of predictive power towards the bias parameter. Thus, GCM's generating mechanism may disregard the

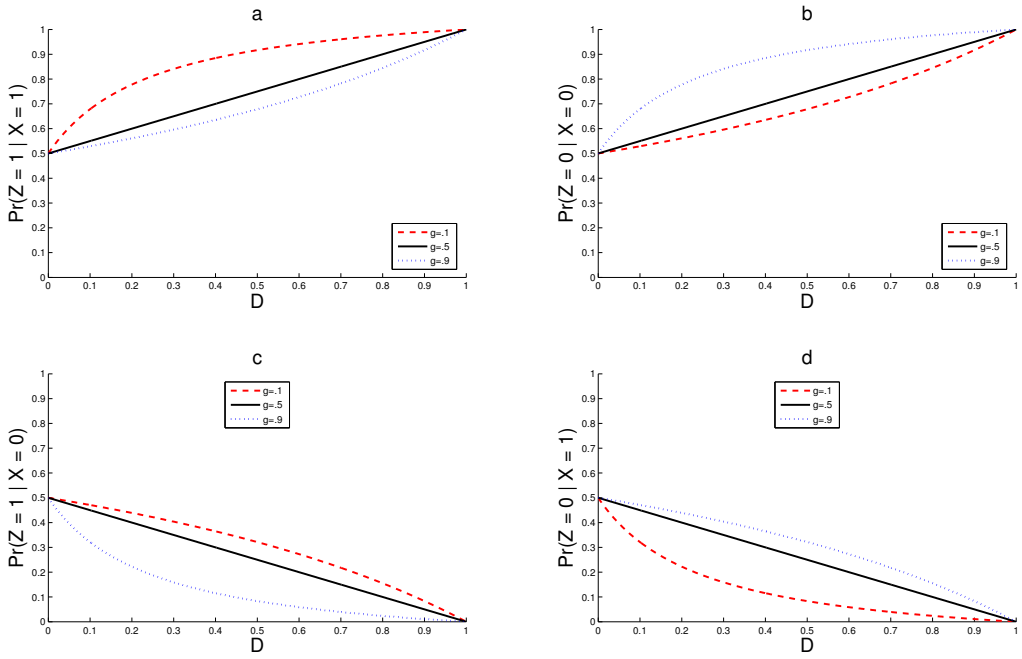

Figure 5.1: The probability of Z, given X (i.e. $P(Z = z|X = x)$) given each parameter of GCM.

influence of expertise in favor of a purely guessing strategy. Basically, the GCM's strategy for constructing an answer key is afflicted with an ABE that obfuscates the influence of expertise.

While the propensity for errors is greatly increased when ABE is not controlled for, most GCM models have either fixed the bias parameter g to .5 or have when Bayesian estimation techniques are applied, the prior probability distribution has been centered around .5 (Batchelder & Ander, 2012). Restricting the distribution of the bias parameter to be near .5 reduces the possibility that any single informant with a low ability would influence the answer key over an informant with greater ability. Such control would restrict the model from preferring an inexperts biased response over an experts unbiased response when constructing the answer key.

An astute observer may find that a strong bias towards an answer would indicate a higher

degree of commitment to one response choice. Thus reducing the number of trials on which a single response from the biased informant may influence the answer key recovery over a competent informant. Consider the probability of an informant responding yes, i.e. $P(X = 1)$. This can be expressed using conditional probabilities, such that:

$$P(X = 1) = \pi P(X = 1|Z = 1) + (1 - \pi)P(X = 1|Z = 0).$$

Substituting the conditional probabilities with parameters found in Axiom 3, we get:

$$P(X = 1) = \pi[D + (1 - D)g] + (1 - \pi)(1 - D)g. \tag{5.6}$$

To see the specific influence of parameter g, we generate a sample of marginal probabilities give a range of parameters. First we set the prior probability that an item is one (i.e. $P(Z = 1)$) at $\pi = .5$ and obtain a set of values between [0,1] for D and g. Figure 2 shows the probability an informant response "Yes," (i.e. P(X=1))

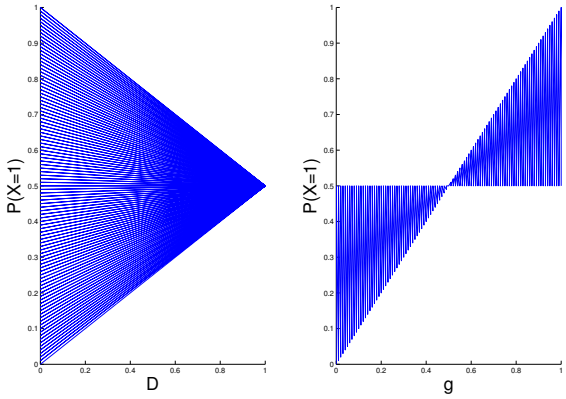

Figure 5.2: The probability of x (i.e. $P(X = 1)$) given each parameter of GCM.

The figure reveals that as an informants ability increases, their response converges to their prior belief probability, regardless of guessing. Although this prior belief drives an informants response, it does so only when $D = 1$, however, less competent informant will rely more on their guessing bias. Therefore, the informants response relies heavily on both their prior belief and their guessing bias. Thus an informants ability, D, acts as an informant regulation system on a guessing bias in favor of their prior belief.

In order to taper of this models dependency on the relative association to an individuals bias we link the distributional assumptions of g to the item in question. Thus if an individual where to rely on guessing, it may do so along with everyone else for each item. Under this new model, the guessing bias is a function of the item. For this to work, without clearly including an unyielding number of parameters to the model, we rely on hierarchical distributions. The following section will describe the new model in more detail.

5.3 Knowledge Gradient Consensus Model

The principle reason for developing the new model, Knowledge Gradient Consensus (KGCM), stems from a desire to represent individual knowledge across a set of items in a more veridical fashion. It was seen in the above section that bias plays an important role in how an informant responds and though it was controlled for in other papers, it seems more appropriate to allow g to represent a systematic bias conditioned on an item. We rely on hierarchical distributions out of concern that the number of parameter may be too great for proper identification.

Axiom 1. (*Cultural Truth*). There exists a single answer key shared by all informants, Z.

Axiom 2. (*Conditional Independence*). The response profile matrix satisfies conditional

independence given by: $Pr[\mathbf{X} = (x_{i,k})_{N\times M}|\mathbf{Z}, \mathbf{H}, \mathbf{F}]$

$$= \prod_{i=1}^{N}\prod_{k=1}^{M} Pr(X_{i,k} = x_{i,k}|Z_k, H_{i,k}, F_{i,k}) \tag{5.7}$$

for all possible realizations $(x_{i,k})$ of the response profile matrix.

Axiom 3. (*Marginal Probabilities*). The marginal probabilities (2) are given by $Pr(X_{i,k} = x_{i,k}|Z_k, H_i, F_i)$

Axiom 4. (*Double High Threshold*). Hit and false alarm parameters for each informant are reparameterized by

$$\forall i, \forall k H_{i,k} = D_{i,k} + (1 - D_{i,k})g_{i,k}, \tag{5.8}$$

$$F_{i,k} = (1 - D_{i,k})g_{i,k}. \tag{5.9}$$

Axiom 5.a. (*Informant Ability and item difficulty*). Each informants ability is a draw from a hierarchical probability distribution centered around $\mu_D(i, k)$ with variance $\sigma_D(i, k)$ for each individual item.

$$D_{i,k} \sim PDF(\mu_D(i, k), \sigma_D(i, k)) \tag{5.10}$$

Axiom 5.b. (*Conditional guessing*). Each informant bias is a draw from a hierarchical

probability distribution function with mean $\mu_g(k)$ and variance $\sigma_g(k)$ for each individual item.

$$g_{i,k} \sim PDF(\mu_g(k), \sigma_g(k)) \tag{5.11}$$

The first 4 axioms of the new model are alike to those in the original GCM; however, we have added the assumption that the model parameters are now dependent on item in the Axiom 5. Axiom 5.1 focuses on the ability parameter and it makes the general assumption that each informant's ability parameter depends on a particular item. This extension of the model is not novel, see Anders and Batchelder 2012; however instead of using a Rasch model like that in Anders and Batchelder's paper, we assume that each ability parameter value is drawn from a joint distribution centered around $mu_{i,k}$ and with a covariance matrix $Sigma_{i,k}$.

Although this dependency is crucial in the new way of conceptualizing the model, it creates an estimation problem due to the limited number of degree's of freedom. In order to alleviate some of the strain of estimating more parameters than degrees of freedom, we set the off-diagonal elements of the covariance matrix to zero. Next, we focus on reducing the number of hyperpriors by instead of having a separate hyperparameter for each item and informant we will only have parameters that focus on relevant groups of each. To do this, we rely on hierarchical distributional methods combined with a latent mixture model (LMM). LMM's are probabilistic models used to represent subpopulations from within an overall population. In our application of LMM's, we use them to identify subgroups within informants and items. By representing the probability distribution of observations in the overall population as a mixture distribution, we can focus on only the parameters responsible for delineating each distribution. This has the effect of reducing the number of free parameters required for the

model to only a few important hyperpriors for each hierarchical probability distributions associated with each subgroup.

We begin by applying the following probability distributions on each parameter:

$Z_t \sim Bernoulli(\pi_k)$

$D_{i,k} \sim Beta(\mu_D[\Gamma(i), \Omega(k)]\tau_D[\Gamma(i), \Omega(k)], (1 - \mu_D[\Gamma(i), \Omega(k)])\tau_D[\Gamma(i), \Omega(k)])$

$g_{i,k} \sim Beta(\mu_g[\Omega(k)]\tau_g[\Omega(k)], (1 - \mu_g[\Omega(k)])\tau_g[\Omega(k)])$

$\Gamma(i) \sim Categorical(\lambda_i)$

$\Omega(k) \sim Categorical(\lambda_k)$

The Beta distribution is commonly used for variables within $[0, 1]$ and since we are interested in the means of each of the distributions, we modify the shape parameters to reflect our needs.

The following hyperpriors were used:

$\pi \sim Beta(1, 1)$

$\mu_D(i, k) \sim dbeta(1, 1)$

$\mu_g(k) \sim dbeta(4, 4)$

$\tau_D(i, k) \sim dgamma(1, 1)$

$\tau_g(k) \sim dgamma(1, 1)$

$\lambda_i \sim Dirichlet(I)$, $I = (1)_{1\times T}$

$\lambda_k \sim Dirichlet(K)$, $I = (1)_{1\times P}$

where T is determined from a scree plot test and P is prespecified.

5.4 Study 1

The first questionnaire was created to study individual differences by focusing on possible areas of knowledge an undergraduate student may have encountered through their academic studies. The goal here is to allow for natural academic interests to yield different expertise among the informants.

5.4.1 Methods

A questionnaire containing 10 questions pertaining to each of the following topics; History, Statistics and English, was administered to participants.

The analysis of the model was conducted using JAGS and the results reported are from a run of 1000 samples with no burnin but with thinning set at 10. Both the standard GCM and modified model are used to analyze the data and their results are reported below.

Since the results are draws from a Bayesian sampler there exists a problem in determining the membership of an individual and only collecting samples from the appropriate distribution without retaining any samples gathered from the other distribution during the sampling procedure. For example, my contribution to CCT is the assumption that an informant does not have the same ability of answering all questions in a questionnaire rather there are some questions an informant is more likely to answer correctly than others. In this case, lets assume there are 3 categories with 10 questions each for which an informant is differentially capable of answering correctly. We can assign for each informant 3 different ability parameters for questions pertaining to each category.

However, in practice we are unaware of the precise number of categories and their proportion of questions in the questionnaire. Therefore we turned to latent mixture models as a possible

solution to finding the membership of each item and in return estimate an informants ability parameters. Again to keep things simple, lets assume there are 3 categories and so we want to estimate 3 different ability parameters for each informant. Now, during the sampling procedure, a value for an informants ability for a given question is taken from one of three distributions depending on the assignment of the question. Naturally, one would expect to obtain the same ability parameter values for all questions associated with one category. Unfortunately since the questions membership is unknown, the informants ability for each question is then a mixture of 3 distributions. So instead of getting the same ability parameter values for all questions within a category we get 10 different ability values for the 10 questions within a category.

Basically the posterior distribution of any given informants ability parameter for a question is a combination of 3 distributions depending on the number of times in the simulation that question was associated with each category. Obviously this is not what we want. So to fix this, we post-process the results based on the indicator variable posterior samples to select out the posterior values of the group-specific parameters.

Now, there are multiple methods to do this, for example, for a set M items, it is possible to categorize the results based on the category variable using the mode assignment for each item individually. Or, we can take the mode assignment of the complete list of items and post-process the results. In our analysis, we found that the two gave the same results.

5.4.2 Informants

N = 29 informants were sampled from Social Science subject pool at the University of California.

	DIC	Deviance
GCM	1100.13452	1074.61226
KGCM	731.581246	621.267123

Table 5.1: Deviance Information Criterion for GCM and KGCM on Experiment 1

5.4.3 Procedures

This questionnaire can be found in the Appendix under chapter 5.

5.4.4 Results 1

Table 5.1 shows the Deviance Information Criterion (DIC) calculated for both the GCM and KGCM.

Table 5.2 displays each item's predicted truth value given by each model along with the ground truth and the average response given by all informants. The values inside the parenthesis are the unrounded probabilities. The averaged response correctly predicted 70 % of the items, GCM correctly predicted 66.66% of the items and KGCM correctly predicted 76.66 % of the items.

5.4.5 Discussion 1

Note, mean responses successfully reconstructed the answer key with 20 correct. This was the same using the GCM model, however, using the new model, we were able to correctly reconstruct the answer key with 23 correct.

GT	Average	GCM	KGCM
1	1 (0.8276)	1 (1)	1 (1)
1	0 (0.4138)	0 (0.031)	0 (0)
1	1 (0.5862)	1 (0.997)	0 (0.1975)
0	1 (0.931)	1 (1)	1 (1)
1	1 (0.6552)	1 (0.999)	1 (0.9797)
1	1 (0.5517)	1 (0.981)	1 (0.7418)
1	1 (0.8276)	1 (1)	1 (0.9949)
1	1 (0.6552)	1 (0.999)	1 (0.7519)
0	1 (0.6207)	1 (0.974)	1 (1)
0	0 (0.4138)	0 (0.172)	0 (0.0608)
1	1 (0.5862)	1 (0.995)	1 (0.7468)
1	0 (0.4483)	0 (0.151)	1 (1)
1	1 (0.5172)	1 (0.856)	1 (1)
1	1 (0.6897)	1 (0.999)	1 (1)
0	0 (0.4828)	1 (0.576)	1 (0.5392)
1	1 (0.6897)	1 (1)	1 (0.719)
1	1 (0.8276)	1 (1)	1 (1)
1	0 (0.4828)	0 (0.413)	1 (0.6101)
0	1 (0.8621)	1 (1)	1 (1)
0	0 (0.0345)	0 (0)	0 (0)
0	1 (0.5862)	1 (0.994)	0 (0.481)
0	1 (0.6552)	1 (0.999)	1 (0.9797)
0	0 (0.1034)	0 (0)	0 (0)
0	0 (0.4138)	0 (0.079)	0 (0)
0	0 (0.2414)	0 (0)	0 (0.0025)
1	1 (0.7241)	1 (1)	1 (0.957)
0	0 (0.1034)	0 (0)	0 (0)
0	1 (0.5517)	1 (0.975)	0 (0.2456)
1	1 (0.6207)	1 (0.998)	1 (1)
1	1 (0.7241)	1 (1)	1 (0.9671)

Table 5.2: Predicted Answers keys by GCM and KGCM where GT = ground truth

5.5 Study 2

The second questionnaire was created to study individual differences pertaining to cultural knowledge The focus here is to find groups of expertise more naturally found in the population. While the last experiment explored information often taught formally, this new experiment, attempts to discover differences in knowledge transmitted through cultural settings.

5.5.1 Methods

The questionnaire created tests knowledge about; film, literature, and sports. Each topic has 10 questions, and three version of this questionnaire was made with three different ordering of the questions to control for

The analysis of the model was conducted using JAGS and the results reported are from a run of 1000 samples with no burnin but with thinning set at 10. Both the standard GCM and modified model are used to analyze the data and their results are reported below.

5.5.2 Informants

80 informants in a psychology class at the University of California, Irvine were administered the questionnaire. Of the 80, 3 did not complete the questionnaire completely, so we omitted their response for the analysis.

5.5.3 Procedures

This questionnaire can be found in the Appendix under chapter 5.

	DIC	Deviance
GCM	2914.85328	2858.81164
KGCM	2277.91716	2039.68858

Table 5.3: Deviance Information Criterion for GCM and KGCM on Experiment 2

5.5.4 Results 2

Table 5.3 shows the DIC results for the GCM and KGCM model.

Table 5.4 shows the average response, GCM, and KGCM answer key predictions along with each unrounded probability in parenthesis. The averaged response correctly predicted 70 % of the items, GCM correctly predicted 73.33% of the items and KGCM correctly predicted 83 % of the items.

5.5.5 Discussion 2

The results provided above have shown that the two models outperform the average, however, it is easy to see that KGCM's predicted answer key approaches the ground truth on three more items than the GCM.

5.6 Conclusion

Cultural Consensus Theory has provided a useful framework for conceptualizing preexisting differences present within a group of people as a means towards scientifically verifiable knowledge. In the past, CCT was almost exclusively used to further an Anthropological hypothetical deductivistic pursuit as a quantifiable tool for theorists looking to employ formalized measures. More recently this theory has broadened its scope, permitting a greater appreciation of the theories applicability across different domains. However, most mathe-

GT	Average	GCM	KGCM
1	0 (0.4935)	1 (0.993)	1 (1)
1	1 (0.6494)	1 (0.998)	1 (1)
0	1 (0.5065)	0 (0.001)	0 (0.3039)
1	1 (0.6753)	1 (1)	1 (1)
1	1 (0.5974)	1 (0.997)	1 (1)
0	0 (0.4156)	0 (0)	0 (0)
1	1 (0.5974)	1 (0.997)	1 (1)
0	1 (0.5325)	1 (0.929)	1 (0.8469)
1	1 (0.6104)	1 (0.997)	1 (1)
1	1 (0.6104)	1 (0.996)	1 (1)
0	0 (0.3117)	0 (0)	0 (0)
0	0 (0.3377)	0 (0)	0 (0)
0	0 (0.4416)	0 (0)	0 (0)
0	0 (0.4675)	0 (0)	0 (0)
1	1 (0.5844)	1 (0.994)	1 (1)
0	0 (0.4935)	0 (0.001)	0 (0.0348)
0	0 (0.4156)	0 (0)	0 (0)
1	1 (0.6494)	1 (0.996)	1 (1)
0	1 (0.6104)	1 (0.998)	1 (1)
1	1 (0.5455)	1 (0.994)	1 (1)
1	1 (0.6883)	1 (1)	1 (1)
0	1 (0.5195)	1 (0.993)	1 (0.8886)
0	1 (0.5714)	1 (0.996)	1 (1)
0	0 (0.4286)	0 (0)	0 (0)
0	1 (0.5325)	1 (0.994)	0 (0.0847)
1	1 (0.5065)	0 (0.498)	1 (1)
1	1 (0.5325)	1 (0.994)	1 (1)
1	0 (0.4805)	0 (0.073)	1 (0.6218)
0	1 (0.5974)	1 (0.995)	1 (1)
1	1 (0.5714)	1 (0.997)	1 (1)

Table 5.4: Predicted Answers keys by GCM and KGCM where GT = ground truth

matical models in connection with this theory have relied solely on an informants expertise to enumerate an unknown answer key. In light of existing research on expertise, (Weiss), leverage provided by ability alone may focus too heavily on incomplete conceptualization of pertinent latent information.

The current paper has further explored the nature of each individual parameter of the most commonly used CCT model, the General Condercet Model. In our analysis of the model parameters, an asymmetric bias effect was shown to negatively influence the construction of an answer key. More specifically, when this effect is not controlled for, model specification errors may lead to a misattribution of parameter influence. In turn, the estimated parameter value of the latent answer key becomes less accurate. Therefore, it is important to for scientific accuracy to modify the models to control

5.7 Reference

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. Psychometrika, 80(1), 151-181.

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. Journal of Mathematical Psychology, 61, 1-13.

Batchelder, W. H., & Romney, A. K. (1989). New results in test theory without an answer key. Mathematical psychology in progress, 229-248.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. Journal of Mathematical Psychology, 56(5), 316-332.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. Psychometrika, 53(1), 71-92.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64(2), 153-168.

Oravecz, Z., Anders, R., & Batchelder, W. H. (2015). Hierarchical Bayesian modeling for test theory without an answer key. Psychometrika, 80(2), 341-364.

Chapter 6

Metric CCT

The goal is to construct a general continuous model of CCT for distance data. Cultural consensus theory (CCT) is an information pooling technique that utilizes natural differences that arise amongst informants to construct a culturally viable answer key to questions with no known answers. One area of otherwise unexplored cultural knowledge using CCT models is that of multidimensional data such as distance predictions. While research on distances between geographical locations has benefited from other methodologies such as multidimensional scaling techniques, it has done so without the help of cognitively based variables that allow us to quantify the degree of knowledge in a collection of informants.

This paper will present a metric response CCT model that will take individual performance abilities along with item difficulty measures into account when constructing distances between objects that satisfy the triangle inequality.

6.1 Introduction

Cultural consensus theory (CCT) is an information pooling technique for reconstructing unknown answers shared by a group of informants. At the basic level, CCT utilizes natural differences that arise amongst informants to construct a culturally viable answer key to questions with no known answers. This is especially useful when clear consensus between informants is otherwise unattainable from aggregation methods such as the majority rule. For example, when there is no clear plurality between alternatives, such that each response alternative has an equal proportion of votes, central measures such as the mean or mode utilized for methods such as the majority rule will not provide a solution. Although CCT is not the only model that will resolve this issue, (e.g. Lee, 2001), it formalizes the solution by weighting each informant's response through axiomatized latent measures such as cognitive ability and systematic bias. These person specific latent variables along with the unknown answer key can be obtained for a variety of scale families.

One otherwise unexplored cultural knowledge with CCT models is that of unbounded psychophysical measures characterized by a complex set of attributes on a continuum shared by informants. Metric based judgments are ubiquitous across the field of psychology where a purely objective quantitative measurement of the degree of similarity between items is unknown. While research on psychological spaces has benefited greatly from both parametric and non-parametric methodologies, it has done so without much help from cognitively based variables shown to quantify the degree of knowledge from a collection of informants and item difficulty. Person specific dimensionality weights have been included to account for individual differences (INDSCAL; Carol and Chang, 1970), however, they have not been used to differentially weight each informant's response with the goal of constructing a culturally accepted stimuli structure.

Psychophysical models come in a variety of forms, such models have been studied on values

of similarity judgments between psychologically motivated objects with known relational attributes through ratings on a scale or triadic comparison (Romney, Brewer, & Batchelder, 1993). These responses are then mapped onto an orthonormal vector space shared by each informant. More recently, MDS type data has become increasingly popular in the field of marketing.

This paper aims at constructing a metric response CCT model that will give the user information about each informant's cognitive measures along with a culturally appropriate view on the structure of the items[1] in a multidimensional space. The reason for working with distances is to study person specific weights used for reconstructing an actual map with known dimensions. Thus providing us with the opportunity to validate the model's parameters as meaningful cognitive variables for known dimensions. After the introduction, the paper will introduce the new model and its properties. Following this, we introduce the data and analyze it using standard individual differences MDS. Afterwards the model is applied and compared to the previous analysis to showcase the increased extraction of information from subject data.

There are two models in question when building a psychophysical model of distances. the first is a spatial model and the second is a distance model. Multidimensional scaling techniques require a certain linear combination of known

6.2 Distance and Spatial model

Generally, when attempting to model psychophysical data one must decide upon a distance model and a spatial model. The first, a distance model, specifically models a subject's stated belief of a distance or similarity between two items. Unless otherwise stated, a distance

[1]Here items is defined as the objects for which an experimenter is interested in, instead of each question asked to the informant.

model provides the necessary framework to predict subject responses to psychophysical data. The second, maps these predicted distances as distances between points in some unknown multidimensional space, known as the spatial model. A choice of the spatial model is left to the researchers discretion so long as it is a formulation of a metric with a corresponding metric space. The two work in concert to yield an appropriate representation of the data.

To date, a common choice for each model has been what Torgerson (1958) calls, deterministic models. These type of models have the distinction that all the variation in the data is accounted for by the informant and item. The models take no further steps to account for systematic error so the viability of the model is left to how well it can approximate the data. Multidimensional scaling techniques such as MDS, INSCAL, etc. all fall under this general framework. In recent years, progress on probabilistic techniques has revitalized the use of psychometric models with the goal of utilizing these techniques for parameter estimation of deterministic models (Michael Lee, Rouder). Although focus has been on using Bayesian assumptions in estimation of the parameters posed by the deterministic model, it requires the modeler to allow for systematic error. In doing so, the models are no longer deterministic but rather, probabilistic.

While this particular distinction between models was first outlined by Torgerson (1958), the problem of misattribution of error between the spatial and distance models was not suspected when probabilistic assumptions are applied to deterministic models. Generally the issue is side swept by redirecting the error so that it falls on the parameters of the spatial model, while ignoring the source of the variation as belonging to the distance estimates.

This has been done with michael lee's and styvers models, possibly also, Rouder. CHeck.

While this relationship has not been fully explored, possibly due to the fact that any distance models can be combined with an ever growing list of spatial models, a clear choice is made in this paper between the spatial and distance model. For mathematical ease, we decide to

use a Euclid's metric system.

6.3 CCT Continuous Model

CCT models developed to account for continuous truth response data commonly assume that an informant draws a latent appraisal[2], Y_{ik}, characterized as deviating from a cultural truth Z_k by some amount determined by ϵ_{ik}. The error random variables are distributed with mean zero and standard deviation σ_{ik}. Furthermore, an assumption is made that each observed response is a function of an informant's competency E_i and item difficulty λ_k thus $\sigma_{ik} = E_i\lambda_k$. Finally, the model includes response bias parameters that account for differences in the observed response from each informants latent appraisal. Thus the observed response, X_{ik}, is a function of two bias parameters that transform the latent appraisal, Y_{ik}, by a scaling bias, a_i and shifting bias b_i, i.e $X_{ik} = a_iY_{ik} + b_i$.

It is rather easy to see that the latent appraisals, Y_{ik}, and shared truth, Z_k, lie on the open interval, $(-\infty, \infty)$, and although the observed responses are not limited to values on the same interval, greater methodological success was reported using Gaussian distributions (Anders & Batchelder, 2014). In fact, it is possible to transform the observed response on the real line to the unit interval using an inverse logit function or from the unit interval to the real line with the logit function. This subtlety allows one to work with a wider range of data sets given the breadth of research with continuous response profiles.

As it stands, the current continuous response model in CCT involves a linear transformation unto $[0, 1]$ where each response in the data set is divided by the largest value. Followed by a transformation unto the $\mathbb{R}$, usually by the logit link function. The first transformation standardizes the distance values to a more manageable range for the second transformation.

[2]Note, the latent appraisal is not observed rather it is an approach adapted from classical test theory (Lord & Novick, 1968).

Unfortunately, it is known that by standardizing values, the dispersion is also adjusted, and thus any information relying on the data variance is lost. The second transformation yields problematic assumptions that cannot be easily reconciled for latent distance approximations. The more important of these problems is the fact that distances are never negative and must satisfy the triangle inequality. Thus these transformations do not help in obtaining an appropriate estimate of true distances since the resulting distribution involves sampling values in $(-\infty, \infty)$. Therefore, any models attempting to find an agreed upon solution between informants that corresponds to the ground truth must make an effort to model the data as it is.

6.4 Metric CCT Model

Assume that each of the N informants provides a continuous estimate on $\mathbb{R}^+$. The random variables, $X_{i,j,j'}$, represents the i^{th} informant's response about two stimuli-objects, where $j, j' \in \Omega$ and $j \neq j'$ and Ω is the set of M stimuli-objects. Alternatively $\mathbf{X} = (X_{i,j,j'})_{N \times M \times M}$ can be viewed as the response profile matrix obtained from each informant. Naturally, the true distance between points is assumed to lie on the positive real line therefore the latent answer key, $Z_{j,j'} \in \mathbb{R}^+$. The resulting symmetric distance matrix, $\mathbf{Z} = (Z_{j,j'})_{M \times M}$ conveniently consolidates the data for the pairwise comparisons between M cities.

Axiom 1 (*Cultural Truth*). There is a single answer key, $\mathbf{Z} = (z_{j,j'})$, applicable to all informants. The distance matrix, $\mathbf{Z}$, is symmetric, non-negative, hollow and satisfies the triangle inequality.

Axiom 2 (*Latent Appraisal*). Each informant draws a latent distance appraisal $\mathbf{Y}$, from

latent coordinate pair $\{x, y\}$, for each city such that, $\mathbf{Y} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} + \epsilon_{i,k}$ satisfies the triangle inequality on a Euclidean plane for each point.

Axiom 3 (*Conditional Independence*). The ϵ_{ik} are mutually independent thus the joint distribution of the latent appraisal is given by:

$$h[(y_{ik}|(Z_k), (\sigma_{ik})] = \prod_i \prod_k f(y_{ik}|Z_k, \sigma_{ik}) \tag{6.1}$$

Axiom 4 (*Precision*). There are knowledge competency parameters $\mathbf{E} = (E_i)_{1\times N}$, where $E_i > 0$. An informants standard appraisal error in the assessment of each is defined as:

$$\sigma_{ij} = E_i \lambda_k \tag{6.2}$$

Axiom 5.a (*Dimension Specific Response Bias*). There are dimension specific bias parameters that act on the latent position of an item.

$$X_{i,j,j'} = a_i Y_{j,j'} = \sqrt{\sum_r^R a_{i,r}(x_{j,r} - x_{j',r})^2} \tag{6.3}$$

Axiom 5.b (*Translation Response Bias*) For each informant, there exists a translation

parameter $\mathbf{B} = (b_i)_{1\times N}$ where $-\infty < b_i < \infty$

$$X_{i,j,j'} = \sqrt{\sum_{r}^{R} a_{i,r}(x_{j,r} - x_{j',r})^2} + b_i \tag{6.4}$$

For every distance response, we are provided with information about the relationship between two stimuli-objects from each participant and usually this is represented by a square symmetric matrix. Axiom 1 of this model generalizes Axiom 1 of (AZB) to include single cultural truths between pairs of objects. This first axiom characterizes the basic assumption of CCT, that there exists an answer-key unknown to the researcher that applies to any informant completing the test/questionnaire. Axiom 2 introduces an error term that is doubly indexed by item and informant applied to the objective answer-key from Axiom 1. Furthermore, while there are other metric functions, we have selected to use Euclid's equation as commonly used in MDS. An advantage of using this equation is that it satisfies the triangle inequality requirement while still allowing us to specify CCT parameters. Axiom 3 is typical for IRT type models that conditions the latent appraisals by the model parameters. Axiom 4 makes the competency parameter be a function of both informant and item. Axiom 5.a and 5.b relates the latent appraisal to the observed response through two linear transformations on the latent truth. The five Axioms delineate the MCCT model in a very similar fashion to the continuous CCT model of (Anders, Zita, & Batchelder, 2013). In fact, if we eliminate the additional, j, j' subscripts on the variables in the four axioms, MCCT reduces to the Continuous CCT model.

Axiom 5.a splits the bias scale parameter so that it can be measured independently for differences dimensions, e.g. North/South or West/East. Albeit the solution does not readily conform to Anders et al. (2014) continuous paper assumptions since that paper suggests a transformation from $[0, \infty)$ to $[0, 1]$ then to $(-\infty, \infty)$ which does not lend itself to a tractable

solution of distances that are positive and that satisfy the triangle inequality.

6.5 Potential Avenues for MCCT

A potential problem for the current model may be found in the bias parameters. One possibility is that if a diffuse prior is assumed for the bias parameters, the shared distances (i.e. Z) as calculated by the euclidean distance equation will not approach the ground truth. The solution is to then assign highly informative priors so that the distances are not underestimated. While this approach has been used by Anders et al. (2014), the problem lies in trying to interpret the results.

In order to understand the problem we take a classical approach of finding maximum likelihood estimators for the parameters in the model. We being by defining the likelihood function as:

$$L(a, b, Z, \lambda|X) = \prod_i \prod_k \frac{1}{\sqrt{2\pi} a_i E_i \lambda_k} exp[\frac{-(x_{ik} - (a_i Z_k + b_i))^2}{2a_i^2 E_i^2 \lambda_k^2}] \tag{6.5}$$

The score function of Z

$$\frac{\partial L}{\partial Z} = -\frac{\sum_i \frac{b_i - x_{ik} + a_i Z_k}{a_i E_i^2 \lambda_k^2}}{\lambda_k^4 \prod_i a_i E_i} exp[\sum_i -\frac{(b_i - x_{ik} - a_i Z_k)^2}{2a_i^2 E_i^2 \lambda_k^2}] \tag{6.6}$$

Setting, $\frac{\partial L}{\partial Z} = 0$, and solving for Z_k we get:

$$\hat{Z}_k = -\frac{\sum_i \frac{b_i - x_{ik}}{a_i E_i^2 \lambda_k^2}}{\sum_i \frac{1}{E_i^2 \lambda_k^2}} \tag{6.7}$$

The score function of b

$$\frac{\partial L}{\partial b} = -\sum_k \frac{b - x_{ik} + aZ_k}{E^2\lambda_k^2 a^2} exp(\sum_k -\frac{(b - x_{ik} + aZ_k)^2}{(2E^2\Lambda_k^2 a^2)}) \tag{6.8}$$

Setting, $\frac{\partial L}{\partial b} = 0$, and solving for b_i we get:

$$\hat{b_i} = \frac{\sum_k \frac{(x_k - aZ_k)}{(E^2\lambda_k^2 a^2)}}{\sum_k (\frac{1}{(E^2\lambda_k^2 a^2)}} \tag{6.9}$$

Setting, $\frac{\partial L}{\partial a} = 0$, and solving for a_i we get:

$$\hat{a_i} = \frac{\sum_k \frac{-b+x}{E^2\lambda^2 T}}{\sum \frac{1}{E^2\lambda^2}} \tag{6.10}$$

CRM assumes that an informant's competency is a measure of how close their estimate fits with the shared truth, Z. After individual bias has been included this assumption might not really hold since it is known that consistency is not necessary nor sufficient for competency (e.g. a person can be consistently wrong) (David Weiss). This suggests that we look elsewhere when allocating the competency parameter to the model, but so far in the current set up of the model the response distribution's variance is a function of competency and item difficulty.

6.6 Conclusion

Work aimed at understanding the parameters and their influence on reconstructing an accepted representation in a multidimensional space is still needed. Most notably, the distinction between a spatial model and a distance model must be made in order to assign the variance appropriately.

This paper has proposed a method for a metric based model, and though more work is needed, the potential benefits of this work may be widespread, as seen through the application of MDS in fields such as marketing and psychology.

6.7 Reference

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. Journal of Mathematical Psychology, 61, 1-13.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 35(3), 283-319.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. Journal of Mathematical Psychology, 45(1), 149-166.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. Psychological Science, 4(1), 28-34.

Torgerson, W. S. (1958). Theory and methods of scaling.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. Psychometrika, 17(4), 401-419.

Appendix A

Appendix Title

A.1 Chapter 2 Appendix

Unequal Variance Signal Detection Hierarchical Model

```
model {

for (i in 1:n) {
    HR[i,1:3] ~ dmulti(h[i,1:3],N)
    FA[i,1:3] ~ dmulti(f[i,1:3],N)
}

for (i in 1:n) {
h[i,1] <- phi(((d[i]/2) - c[i,1])/tau[i])
h[i,2] <- phi(((d[i]/2)-c[i,2])/tau[i])-phi(((d[i]/2)-c[i,1])/tau[i])
h[i,3] <- phi((c[i,2] - (d[i]/2))/tau[i])

f[i,1] <- phi((-d[i]/2) - c[i,1])
f[i,2] <- phi((-d[i]/2) - c[i,2]) - phi((-d[i]/2)-c[i,1])
f[i,3] <- phi((d[i]/2) + c[i,2])
}

for (i in 1:n) {
c[i,1] ~ dnorm(muc1,sigmac1)T(c[i,2],)
```

```
c[i,2] ~ dnorm(muc2,sigmac2)
    d[i] ~ dgamma(mud^2/sigmad, mud/sigmad)
    tau[i] <- 1/tmp[i] +1
    tmp[i] ~ dgamma(mut, 1/sigmat)
}

# dprime hyperpriors

mud ~ dgamma(1,1)
    #sigmad ~ dgamma(1,1)


lambdad ~ dgamma(1,1)
    sigmad <- 1/sqrt(lambdad)

# sigma hyperpriors
mut ~ dgamma(1,1)
lambdat ~ dgamma(1,1)
    sigmat <- 1/sqrt(lambdat)
#sigmat ~ dgamma(1,1)

# 1st criterion hyperpriors
muc1 ~ dnorm(0,1)
lambdac1 ~ dgamma(1,1)
    sigmac1 <- 1/sqrt(lambdac1)

# 2st criterion hyperpriors
    muc2 ~ dnorm(0,1)
lambdac2 ~ dgamma(1,1)
    sigmac2 <- 1/sqrt(lambdac2)
}
```

Double High Threshold Hierarchical Model

```
model {
for (i in 1:n) {
    HR[i,1:3] ~ dmulti(h[i,1:3],N)
    FA[i,1:3] ~ dmulti(f[i,1:3],N)
}
```

```
for (i in 1:n) {
    h[i,1] <- theta[i,1] + (1-theta[i,1]) * alpha[i] * g[i]
    h[i,2] <- (1-theta[i,1]) * (1-alpha[i])
    h[i,3] <- (1-theta[i,1])  * alpha[i] * (1-g[i])

    f[i,1] <- (1-theta[i,2]) * alpha[i] * g[i]
    f[i,2] <- (1-theta[i,2]) * (1-alpha[i])
    f[i,3] <- theta[i,2] + (1-theta[i,2]) * alpha[i] * (1-g[i])
}

for (i in 1:n) {
# Probit transformation
       for(jk in 1:P){
          logit(theta[i,jk]) <- theta.probit[i,jk]
          theta.probit[i,jk] <- mu[jk] + xi[jk]*delta[i,jk]
       }

       # Prior for unscaled participant effects
       delta[i,1:P] ~ dmnorm(mudelta[1:P],Tprec[1:P,1:P])

     logit(g[i]) <- pg[i]
     pg[i] ~ dnorm(meanpg,precpg)

     logit(alpha[i]) <- palpha[i]
     palpha[i] ~ dnorm(meanpalpha,precpalpha)
}

# Hyperpriors

   for(jk in 1:P){
      mudelta[jk] <- 0
      mu[jk] ~ dnorm(0,1)
      xi[jk] ~ dunif(0,100)
   }

   Tprec[1:P,1:P] ~ dwish(W[1:P,1:P],df)
   df <- P+1

   T[1:P,1:P] <- inverse(Tprec[1:P,1:P])

  # Scale sigma's and compute parameter correlations
    for(jk in 1:P){
       for(prme in 1:P){
       # Off-diagonal elements of S
          rho[jk,prme] <- T[jk,prme]/sqrt(T[jk,jk]*T[prme,prme])
```

```
       }
# Diagonal elements of S
       sigma[jk] <- xi[jk]*sqrt(T[jk,jk])
    }

meanpg ~ dnorm(0,.1)
meanpalpha ~ dnorm(0,.1)

SSA ~ dgamma(1,1)
precpalpha <- 1/sqrt(SSA)

SSG ~ dgamma(1,1)
precpg <- 1/sqrt(SSG)

}
```

A.2 Chapter 3 Appendix

The equations for the sixteen response patterns given the model are a sum of products over the parameters of the model. This is because the model fits into the class of Multinomial Processing Tree (MPT) models (Batchelder et al 1997) and a defining property of MPT models is that the expressions for the sixteen category probabilities, for any word, are a sum of products over the parameters of the model (Batchelder & Riefer, 1999; Hu and Batchelder, 1994). In summary the sum of products over the parameters of the model is composed of enumerated state sequences and their associated retrieval parameters for a given response pattern. The thirteen state sequences are: UUUU, UUII, UUIL, UULL, UIII, UIIL, UILL, ULLL, IIII, IIIL, IILL, ILLL, and LLLL. The equations for the sixteen response patterns are obtained by first enumerating the thirteen possible sequences of states over the four trials. For example, the state sequence UUII is represented by the transition probabilities: $(1 - r\mathrm{x})(1 - a\mathrm{x})(1 - r\mathrm{y})(1 - a\mathrm{y})(1 - r\mathrm{z})a\mathrm{z}\ [\ (1 - t\mathrm{z}) + t\mathrm{z}(1 - bi)]$. The parameters are indexed by x, y, and z to illustrate the fact that the parameters used to calculate the probability of a response sequence, given the model for a word, depend on where that word is in the study list for each study trial. The product of the terms inside the first four parenthesis represents the probability the word remains in the U-State for the first two study trials, and the remaining terms give the probability that the word transitions to the I-State on the third study trial and then fails to make the transition to the L-State on the third test trial.

Each one of these thirteen state sequences is associated with several possible response patterns. For example, the state sequence UULL can result in the response sequences 0000, 0001, 0010, 0011, with recall probabilities, $[(1 - l1{,}z)(1 - l2)]$, $[(1 - l1{,}z)l2]$, $[l1{,}z(1 - l2)]$, $[l1{,}zl2]$, respectively. For a given response sequence, each of the thirteen enumerated state sequences generates a probability of that sequence and these are summed to give the probability of that response sequence in terms of the model's parameters. Note there are instances where not all of the thirteen state sequences can produce a particular observed

response pattern, for example the sequence UULL cannot result in the response sequence 1111 and there are also instances where some state sequences are repeated such as UULL occurs twice in the response sequence 0011. In such cases, the sum includes only the appropriate state patterns. The likelihood function (e.g Riefer & Batchelder 1988) for the model is a product-multinomial distribution, with a sixteen term multinomial for each of the ten words with category probabilities represented by the expression corresponding to the possible response sequences. The likelihood function gives the probability of the obtained data as a function of variations in the parameters of the model, and it is a necessary component of various classical and Bayesian inference procedures used to estimate the parameters of a model.

```
model {
for (i in 1:N) {
   for (j in 1:Subj) {
     XX[i,1:16,j] ~ dmulti(PP[i,1:16,j],1)
}}
# --------------Posterior Predictive ---------------------------------
for (i in 1:N) {
   for (j in 1:Subj) {
     PPXX[i,1:16,j] ~ dmulti(PP[i,1:16,j],1)
}}
# -------------- Pr(Data | Model ) ---------------------------------
for (i in 1:N) {
   for (j in 1:Subj) {
```

$$\text{PP[i,1,j]} <- (1-l1[x[i],j])*(1-l1[y[i],j])*(1-l1[z[i],j])*(1-l2[j])*r[x[i],j] + (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*(1-l1[y[i],j])*(1-l1[z[i],j])*(1-l2[j]) + (1-r[x[i],j])*$$

$$a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*v[z[i],j]*(1-l1[z[i],j])*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*(1-l1[y[i],j])*(1-l1[z[i],j])*(1-l2[j])+\ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*v[z[i],j]*(1-l1[z[i],j])*(1-l2[j])+\ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*(1-v[z[i],j])*(1-t[z[i],j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*(1-r[z[i],j])*(1-a[z[i],j])+\ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*(1-r[z[i],j])*a[z[i],j]*(1-t[z[i],j])+(1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*(1-v[z[i],j])*(1-t[z[i],j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*r[z[i],j]*(1-l1[z[i],j])*(1-l2[j])$$

$$\text{PP[i,2,j]} <-\ (1-l1[x[i],j])*(1-l1[y[i],j])*l1[z[i],j]*(1-l2[j])*r[x[i],j]+\ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*(1-l1[y[i],j])*l1[z[i],j]*(1-l2[j])+\ 1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*v[z[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*(1-l1[y[i],j])*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*v[z[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*v[z[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*(1-v[z[i],j])*t[z[i],j]*(1-b[i,j])+\ 1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*(1-r[z[i],j])*a[z[i],j]*t[z[i],j]*(1-b[i,j])+(1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*(1-v[z[i],j])*t[z[i],j]*(1-b[i,j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*r[z[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*(1-r[z[i],j])*a[z[i],j]*t[z[i],j]*b[i,j]*(1-l2[j])$$

$$\text{PP[i,3,j]} <-\ (1-l1[x[i],j])*l1[y[i],j]*(1-l1[z[i],j])*(1-l2[j])*r[x[i],j]+\ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*l1[y[i],j]*(1-l1[z[i],j])*(1-l2[j])+\ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*b[i,j]*(1-l1[z[i],j])*(1-l2[j])+(1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*v[z[i],j]*(1-l1[z[i],j])*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*l1[y[i],j]*(1-l1[z[i],j])*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*b[i,j]*(1-l1[z[i],j])*(1-l2[j])+\ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*(1-b[i,j])*v[z[i],j]*(1-l1[z[i],j])*(1-l2[j])+\ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-v[z[i],j])*(1-t[z[i],j])+\ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*(1-v[z[i],j])*(1-t[z[i],j])$$

$$\text{PP[i,4,j]} <-\ (1-l1[x[i],j])*l1[y[i],j]*l1[z[i],j]*(1-l2[j])*r[x[i],j]+\ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*l1[y[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*b[i,j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*v[z[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*l1[y[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*b[i,j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*(1-b[i,j])*v[z[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*(1-b[i,j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*r[z[i],j]*l1[z[i],j]*(1-l2[j])+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*(1-r[z[i],j])*a[z[i],j]*t[z[i],j]*b[i,j]*(1-l2[j])+(1-r[x[i],j])*a[x[i],j]*$$

$(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*(1-v[z[i],j])*t[z[i],j]*(1-b[i,j])$

PP[i,5,j] <- $(1-l1[x[i],j])*(1-l1[y[i],j])*(1-l1[z[i],j])*l2[j]*r[x[i],j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*(1-l1[y[i],j])*(1-l1[z[i],j])*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*v[z[i],j]*(1-l1[z[i],j])*l2[j]+(1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*(1-l1[y[i],j])*(1-l1[z[i],j])*l2[j]+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*v[z[i],j]*(1-l1[z[i],j])*l2[j]+(1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*r[z[i],j]*(1-l1[z[i],j])*l2[j]$

PP[i,6,j] <- $(1-l1[x[i],j])*(1-l1[y[i],j])*l1[z[i],j]*l2[j]*r[x[i],j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*(1-l1[y[i],j])*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*v[z[i],j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*(1-t[y[i],j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*(1-l1[y[i],j])*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*v[z[i],j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*(1-t[y[i],j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*r[z[i],j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*(1-a[y[i],j])*(1-r[z[i],j])*a[z[i],j]*t[z[i],j]*b[i,j]*l2[j]$

PP[i,7,j] <- $(1-l1[x[i],j])*l1[y[i],j]*(1-l1[z[i],j])*l2[j]*r[x[i],j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*l1[y[i],j]*(1-l1[z[i],j])*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*b[i,j]*(1-l1[z[i],j])*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*v[z[i],j]*(1-l1[z[i],j])*l2[j]+(1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*l1[y[i],j]*(1-l1[z[i],j])*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*b[i,j]*(1-l1[z[i],j])*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*(1-b[i,j])*v[z[i],j]*(1-l1[z[i],j])*l2[j]$

PP[i,8,j] <- $(1-l1[x[i],j])*l1[y[i],j]*l1[z[i],j]*l2[j]*r[x[i],j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*v[y[i],j]*l1[y[i],j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*b[i,j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*v[z[i],j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*a[x[i],j]*(1-t[x[i],j])*(1-v[y[i],j])*t[y[i],j]*(1-b[i,j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*r[y[i],j]*l1[y[i],j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*b[i,j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*(1-b[i,j])*v[z[i],j]*l1[z[i],j]*l2[j]+ \ (1-r[x[i],j])*(1-a[x[i],j])*(1-r[y[i],j])*a[y[i],j]*t[y[i],j]*(1-b[i,j])*(1-v[z[i],j])*t[z[i],j]*b[i,j]*l2[j]$

PP[i,9,j] <- $l1[x[i],j]*(1-l1[y[i],j])*(1-l1[z[i],j])*(1-l2[j])*r[x[i],j]+ \ (1-r[x[i],j])*a[x[i],j]*t[x[i],j]*b[i,j]*(1-l1[y[i],j])*(1-l1[z[i],j])*(1-l2[j])+ \ (1-r[x[i],j])*a[x[i],j]*t[x[i],j]*(1-b[i,j])*v[y[i],j]*(1-l1[y[i],j])*(1-l1[z[i],j])*(1-l2[j])+(1-r[x[i],j])*a[x[i],j]*t[x[i],j]*(1-b[i,j])*(1-v[y[i],j])*(1-t[y[i],j])*v[z[i],j]*(1-l1[z[i],j])*(1-l2[j])+ \ (1-r[x[i],j])*a[x[i],j]*t[x[i],j]*(1-b[i,j])*(1-v[y[i],j])*(1-t[y[i],j])*(1-v[z[i],j])*(1-t[z[i],j])$

PP[i,10,j] <- $l1[x[i],j]*l1[y[i],j]*(1-l1[z[i],j])*(1-l2[j])*r[x[i],j]+ \ (1-r[x[i],j])*a[x[i],j]*t[x[i],j]*b[i,j]*l1[y[i],j]*(1-l1[z[i],j])*(1-l2[j])+ \ (1-r[x[i],j])*a[x[i],j]*$

$t[x[i],j] * (1 - b[i,j]) * v[y[i],j] * l1[y[i],j] * (1 - l1[z[i],j]) * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * b[i,j] * (1 - l1[z[i],j]) * (1 - l2[j]) + (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * v[z[i],j] * (1 - l1[z[i],j]) * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * (1 - v[z[i],j]) * t[z[i],j] * (1 - b[i,j]) * (1 - t[z[i],j])$

PP[i,11,j] <- $l1[x[i],j] * (1 - l1[y[i],j]) * l1[z[i],j] * (1 - l2[j]) * r[x[i],j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * b[i,j] * (1 - l1[y[i],j]) * l1[z[i],j] * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * v[y[i],j] * (1 - l1[y[i],j]) * l1[z[i],j] * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * (1 - t[y[i],j]) * v[z[i],j] * l1[z[i],j] * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * (1 - t[y[i],j]) * (1 - v[z[i],j]) * t[z[i],j] * b[i,j] * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * (1 - t[y[i],j]) * (1 - v[z[i],j]) * t[z[i],j] * (1 - b[i,j])$

PP[i,12,j] <- $l1[x[i],j] * l1[y[i],j] * l1[z[i],j] * (1 - l2[j]) * r[x[i],j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * b[i,j] * l1[y[i],j] * l1[z[i],j] * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * v[y[i],j] * l1[y[i],j] * l1[z[i],j] * (1 - l2[j]) + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * b[i,j] * l1[z[i],j] * (1 - l2[j]) + (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * v[z[i],j] * l1[z[i],j] * (1 - l2[j]) + (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * (1 - v[z[i],j]) * t[z[i],j] * b[i,j] * (1 - l2[j]) + (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * (1 - v[z[i],j]) * t[z[i],j] * (1 - b[i,j])$

PP[i,13,j] <- $l1[x[i],j] * (1 - l1[y[i],j]) * (1 - l1[z[i],j]) * l2[j] * r[x[i],j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * b[i,j] * (1 - l1[y[i],j]) * (1 - l1[z[i],j]) * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * v[y[i],j] * (1 - l1[y[i],j]) * (1 - l1[z[i],j]) * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * (1 - t[y[i],j]) * v[z[i],j] * (1 - l1[z[i],j]) * l2[j]$

PP[i,14,j] <- $l1[x[i],j] * l1[y[i],j] * (1 - l1[z[i],j]) * l2[j] * r[x[i],j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * b[i,j] * l1[y[i],j] * (1 - l1[z[i],j]) * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * v[y[i],j] * l1[y[i],j] * (1 - l1[z[i],j]) * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * b[i,j] * (1 - l1[z[i],j]) * l2[j] + (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * v[z[i],j] * (1 - l1[z[i],j]) * l2[j]$

PP[i,15,j] <- $l1[x[i],j] * (1 - l1[y[i],j]) * l1[z[i],j] * l2[j] * r[x[i],j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * b[i,j] * (1 - l1[y[i],j]) * l1[z[i],j] * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * v[y[i],j] * (1 - l1[y[i],j]) * l1[z[i],j] * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * (1 - t[y[i],j]) * v[z[i],j] * l1[z[i],j] * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * (1 - t[y[i],j]) * (1 - v[z[i],j]) * t[z[i],j] * b[i,j] * l2[j]$

PP[i,16,j] <- $l1[x[i],j] * l1[y[i],j] * l1[z[i],j] * l2[j] * r[x[i],j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * b[i,j] * l1[y[i],j] * l1[z[i],j] * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * v[y[i],j] * l1[y[i],j] * l1[z[i],j] * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * b[i,j] * l1[z[i],j] * l2[j] + (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * v[z[i],j] * l1[z[i],j] * l2[j] + \quad (1 - r[x[i],j]) * a[x[i],j] * t[x[i],j] * (1 - b[i,j]) * (1 - v[y[i],j]) * t[y[i],j] * (1 - b[i,j]) * (1 - v[z[i],j]) * t[z[i],j] * b[i,j] * l2[j]$

```
}}
# ---- Hierarchical Parameter Distributions With Order Statistics ----------
for (i in 1:N) {
  for (j in 1:Subj) {
    r[i,j] <- r3[11-i,j]
}}
for (j in 1:Subj) {
  r3[1:10,j] <- sort(r2[1:10,j])
  t[1:10,j]  <- sort(t2[1:10,j])
}
for (i in 1:N) {
  for (j in 1:Subj) {
        probit(a[i,j])  <- AA[i,j]
    probit(b[i,j])  <- BB[i,j]
    probit(r2[i,j]) <- r1[i,j]
    probit(t2[i,j]) <- t1[i,j]
     probit(l1[i,j]) <- LL1[j]
    probit(v[i,j])  <- VV[i,j]
}}
for (j in 1:Subj) {
  probit(l2[j]) <- LL2[j]
}
for (i in 1:N) {
  for (j in 1:Subj) {
    AA[i,j] ~ dnorm(muAA[i], sigmaAA[i])
```

```
    BB[i,j] ~ dnorm(muBB[i], sigmaBB[i])
    r1[i,j] ~ dnorm(muRR[i], sigmaRR[i])
    t1[i,j] ~ dnorm(muTT[i], sigmaTT[i])
       VV[i,j] ~ dnorm(muVV[i], sigmaVV[i])
}}
for (j in 1:Subj) {
  LL1[j] ~ dnorm(muL1,sigmaL1)
  LL2[j] ~ dnorm(muL2,sigmaL2)
}
for (i in 1:N) {
  muAA[i] ~ dnorm(0,1)
  muBB[i] ~ dnorm(0,1)
  muRR[i] ~ dnorm(0,1)
  muTT[i] ~ dnorm(0,1)
  muVV[i] ~ dnorm(0,1)
  lambdaAA[i] ~ dgamma(5,5)
  lambdaBB[i] ~ dgamma(5,5)
  lambdaRR[i] ~ dgamma(5,5)
  lambdaTT[i] ~ dgamma(5,5)
  lambdaVV[i] ~ dgamma(5,5)
  sigmaAA[i] <- 1/sqrt(lambdaAA[i])
  sigmaBB[i] <- 1/sqrt(lambdaBB[i])
  sigmaRR[i] <- 1/sqrt(lambdaRR[i])
  sigmaTT[i] <- 1/sqrt(lambdaTT[i])
  sigmaVV[i] <- 1/sqrt(lambdaVV[i])
```

```
}
muL1 ~ dnorm(0,1)
lambdaL1 ~ dgamma(5,5)
sigmaL1 <- 1/sqrt(lambdaL1)
muL2  ~ dnorm(0,1)
lambdaL2 ~ dgamma(5,5)
sigmaL2 <- 1/sqrt(lambdaL2)
}
```

A.3 Chapter 4 Appendix

The expanded model maintains the same rank as the model in the smaller space. Since the transformation does not change the model it is possible to go from the expanded model back to the original model.

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} l & (1-l) & 0 & 0 & 0 \\ 0 & 0 & t & (1-t) & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{T} = \mathbf{\Sigma}\tilde{\mathbf{T}}\mathbf{\Gamma}^{-1}\mathbf{\Lambda}^{-1} \tag{A.1}$$

Proof:

$$\mathbf{T} = \mathbf{\Sigma}\mathbf{\Lambda}^{-1}\mathbf{T}\mathbf{\Sigma}\mathbf{\Gamma}\mathbf{\Gamma}^{-1}\mathbf{\Lambda} \tag{A.2}$$

since $\mathbf{\Sigma}\mathbf{\Lambda}^{-1} = \mathbf{I}$. Where $\mathbf{I}$ is the identity matrix.

$$\mathbf{T} = \mathbf{I}\mathbf{T}\mathbf{\Sigma}\mathbf{I}\mathbf{\Lambda} \tag{A.3}$$

$$\mathbf{T} = \mathbf{I}\mathbf{T}\mathbf{I}\mathbf{I} \tag{A.4}$$

$$\mathbf{T} = \mathbf{T} \tag{A.5}$$

A.4 Chapter 5 Appendix

A.4.1 Questionnaire 1

English ; History ; Statistics

True/False Questions

1 The Northwest Ordinance laid out the requirements for western territories to become states.

2 The binominal distribution assumes that the random variable is the result of counting.

3 The main verb and the direct object are not normally separated in a sentence.

4 When the fighting began during the American Revolution, most Americans wanted the colonies to be independent from Great Britain.

5 Adverbs can modify adjectives.

6 Some adjectives end with -ly.

7 During the first half of the nineteenth century, the United States grew more rapidly in population than Britain or Europe.

8 We normally use an object pronoun after a preposition.

9 The graph of a discrete distribution is a smooth curve.

10 Two events that are mutually exclusive events, are also complements of each other.

11 Artisans, displaced by the factory system, formed the first American labor unions.

12 If two nonempty sets are independent, they cannot be disjoint.

13 In the newly created states, the privileges that churches enjoyed in the colonial era were largely stripped away.

14 Those elements are not in Set A are called the complement of A.

15 The set of all basic outcomes of an experiment is called the union of the experimental set.

16 The mean, median and the mode will all be equal when the distribution is symmetric.

17 The sample standard deviation is a point estimate for the population standard deviation

18 Many adjectives ending -ible/able can come either before or after a noun

19 The number of cars that go passed the drive-in window at a local bank each hour is an example of a continuous random variable.

20 The past tense of "must" is "musted".

21 Questions always use an auxiliary verb.

22 England's first experience with colonization was in Virginia.

23 The median is the value that occurs most often in a sample data.

24 As a result of the Treaty of Paris of 1783, the new American nation's westward boundary was the Blue Ridge Mountains.

25 Native Americans were pleased with the outcome of the Revolution because it reduced the desire of colonists for western land.

26 An attributive adjective comes before a noun.

27 "Used to doing" and "used to do" mean approximately the same thing.

28 The shortest possible sentence contains a subject, a verb and an object.

29 The first Europeans to settle in the Hudson River Valley were the Dutch.

30 The English Reformation began with a political dispute between king and pope not with a religious dispute over matters of theology.

A.4.2 Questionnaire 2

Literature ; Sports ; Film T/F Question

1 Director Tim Burton frequently collaborates with composer Danny Elfman, such as in 1985's Pee-wee's Big Adventure and 1988's Beetlejuice.

2 Steven King is famous for writing horror novels

3 The long sword is a type of sword used in fencing contests

4 To Kill a Mocking Bird was a novel written by Harper Lee

5 Charles Dickens wrote the novel titled A Tale of Two Cities

6 Mary Shelley wrote the horror novel Dracula

7 The Good, the Bad, and the Ugly (1966), Dirty Harry (1971), and Gran Torino (2008) all

star Clint Eastwood.

8 Martin Scorsese directed The Godfather (1972).

9 The Golden State Warriors won the National Basketball Association championship in 2015

10 Star Wars: The Force Awakens (2015) currently holds the record for the biggest opening weekend at the domestic box office.

11 In Agatha Christies series of detective novels, the start detective was Mike Hammer

12 The Lord of the Rings was written by Robert Jordan

13 In racquetball only the receiver of a serve can win a point

14 The Martian (2015) won "Best Picture" at the Academy Awards this year.

15 Germany was the winner of the World Cup Soccer Championship in 2014

16 Casablanca (1942) takes place during the American Civil War.

17 A professional baseball game can end in a tie

18 Alfred Hitchcock is famous for directing psychological thrillers, such as Rear Window (1954), Vertigo (1958), and Psycho (1960).

19 Nathaniel Hawthorne was a famous 19th century novelist from England

20 A safety in American football is a scoring play worth 2 points

21 The Indianapolis 500 is an annual auto race

22 Ping Pong is a sport contested in the Summer Olympics

23 The Wizard of Oz (1939) was the first film in full color.

24 Ezra Pound wrote the famous poem titled The Love Song of J. Alfred Prufrock

25 The curveball and cannon blast are names for types of pitches in baseball

26 The novel, Madame Bovery was written by Gustave Flaubert

27 Robert Louis Stevenson wrote the novel Treasure Island

28 Beauty and the Beast (1991) was the first animated feature film to be nominated for "Best Picture" at the Academy Awards.

29 The Friday the 13th horror film franchise features Freddy Krueger.

30 The world record in the high jump is over 8 feet.

A.4.3 Model Code

```
model {

for (i in 1:N) {
    for (k in 1:M){
        pY[i,k] <- (D[i,k] * z[k]) + ((1-D[i,k]) * g[i,k])
        Y[i,k] ~ dbern(pY[i,k])
    }
}

for (i in 1:N){
  for (k in 1:M){
    D[i,k] ~ dbeta(dmu[e2[i],e[k]]*dth[e2[i],e[k]],(1-dmu[e2[i],e[k]])*dth[e2[i],e[k]])
    g[i,k] ~ dbeta(gmu[e[k]]*gth[e[k]],(1-gmu[e[k]])*gth[e[k]])
   }
}

# -------- Answer Key Parameters ----------------------------------------
for (k in 1:M){
    z[k] ~ dbern(p[k])
    p[k] ~ dbeta(1,1)
}

# -------- Classifies the different clusters of items/informants -------

for (k in 1:M){
e[k] ~ dcat(pe[])
}

for (i in 1:N){
    e2[i] ~ dcat(pe2[])
}

pe[1:P] ~ ddirch(alpha)
pe2[1:T] ~ ddirch(alpha2)


# -------- Hierarchical Hyperpriors -------------------------------------

 for (j in 1:T){
```

```
        gmu[j] ~ dbeta(4,4)             # <- centered around .5
        gth[j] ~ dgamma(1,1)            # <- Diffuse Gamma

for (k in 1:P){
        dmu[j,k]  ~ dbeta(1,1)          # <- Uniform Distribution
        dth[j,k] ~ dgamma(1,1)          # <- Diffuse Gamma
    }
 }

for (j in 1:P){
    alpha[j] <- 1
}

for (j in 1:T){
    alpha2[j] <- 1
}

}
```