ASSESSING SINGLE- AND DUAL-PROCESS ACCOUNTS OF RECOGNITION MEMORY USING HIERARCHICAL BAYESIAN MODELS

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

MICHAEL S. PRATTE

Dr. Jeffrey N. Rouder, Dissertation Supervisor
MAY 2010
The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

ASSESSING SINGLE- AND DUAL-PROCESS ACCOUNTS OF RECOGNITION MEMORY USING HIERARCHICAL BAYESIAN MODELS

presented by Michael Pratte,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

________________________________________

Dr. Jeffrey Rouder

________________________________________

Dr. Moshe Naveh-Benjamin

________________________________________

Dr. Michael Stadler

________________________________________

Dr. Paul Speckman

________________________________________

Dr. Douglas Steinley
ACKNOWLEDGMENTS

• Dr. Jeffrey Rouder, for teaching me or generally supporting my education in just about anything remotely related to psychology, and many things that were not.

• Dr. Richard Morey, for providing me with the knowledge, mentality, and coffee breaks that were vital to my success in graduate school.

• My parents, for allowing me to build, take apart, or burn just about anything for the sake of learning about the world.

• My wife Sarah, for teaching me that there are things in life to be enjoyed besides school and sleep.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS .................................................. ii

LIST OF TABLES ....................................................... vii

LIST OF FIGURES ..................................................... viii

ABSTRACT ............................................................ x

CHAPTER ............................................................... 1

1 Introduction ......................................................... 1

  1.1 Averaging Data In Non-linear Models ......................... 2

  1.1.1 An Example of Averaging in Learning ..................... 4

  1.2 Recognition Memory Data ...................................... 5

    1.2.1 Old-New Recognition .................................. 5

    1.2.2 Receiver Operating Characteristic Curves ............. 7

    1.2.3 Confidence Ratings ................................... 7

  1.3 Models of Recognition Memory ................................ 9

    1.3.1 Equal Variance Signal Detection ...................... 9

    1.3.2 Unequal-Variance Signal Detection .................. 12

    1.3.3 Yonelinas’ Dual-Process Model ...................... 14

  1.4 Separating Process and Nuisance Variability ............... 17

  1.5 The Effects of Averaging In Recognition Memory ........... 18

2 Accounting for Participant and Item Effects in UVSD ........... 24

  2.1 Hierarchial UVSD Model Specification ....................... 25

  2.2 Model Analysis ............................................... 29

    2.2.1 An Introduction to Bayesian Analysis ................ 30
2.2.2 Priors for UVSD Model ........................................ 37
2.2.3 Conditional Posterior Distributions ......................... 38
2.2.4 Sampling Algorithms ......................................... 41
2.2.5 Model Performance .......................................... 45

2.3 Model Restrictions and Comparisons .............................. 47

2.4 Experiment 1 ..................................................... 49
2.4.1 Method ....................................................... 50

2.5 Model Fits ........................................................ 51
2.5.1 Measuring Variance Components .............................. 51
2.5.2 The Structure of Lag, Participant, and Item Effects .......... 54

2.6 Interpreting $\sigma^2$ ............................................ 57

3 Hierarchical Dual- and Single-Process Models of Recognition Memory .................................................. 61
3.1 Extended Hierarchical UVSD Model ............................... 63
3.1.1 Model Specification ........................................... 63
3.1.2 Model Analysis ............................................... 64
3.1.3 Model Performance ......................................... 66
3.1.4 E-UVSD Fit to Experiment 1 .................................. 66

3.2 Gamma Signal Detection Model .................................. 70
3.2.1 Model Specification ........................................... 71
3.2.2 Model Analysis ............................................... 73

3.3 Hierarchical Dual-Process Model ................................ 74
3.3.1 Model Specification ........................................... 74
3.3.2 Model Analysis ............................................... 75
3.3.3 Model Performance ......................................... 77
4 Selective Influence of Dual-Process Parameters

4.1 Previous Claims of Selective Influence
  4.1.1 Reliance on Averaged Data
  4.1.2 Reliance on Remember-Know Paradigm
  4.1.3 Lack of Selective Influence in Confidence Ratings
  4.1.4 Small Effects

4.2 Experiment 2: Biasing Confidence Ratings
  4.2.1 Methods
  4.2.2 Results
  4.2.3 Discussion

4.3 Experiment 3: LOP & Match-Mismatch
  4.3.1 Methods
  4.3.2 Results
  4.3.3 Discussion

4.4 Experiments 4a-c: Increasing the Match-Mismatch Effect
  4.4.1 Experiment 4a: Modality Match-Mismatch
  4.4.2 Experiment 4b: Fast Study Presentation
  4.4.3 Experiment 4c: Shorter Study-Test Lists
  4.4.4 Discussion

4.5 Experiment 5: Response Deadline; Long Study List
  4.5.1 Methods
  4.5.2 Results
4.5.3 Discussion .................................................. 119

4.6 Experiment 6: Response Deadline; Short Study List .............. 121
  4.6.1 Methods ................................................... 121
  4.6.2 Results ................................................... 122

4.7 Discussion ..................................................... 127

5 Summary and Concluding Remarks ................................. 130

BIBLIOGRAPHY .................................................. 134

References .......................................................... 134

VITA ............................................................... 143
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Model Selection Statistics</td>
<td>83</td>
</tr>
<tr>
<td>4.1 Results From Gregg &amp; Gardiner</td>
<td>95</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The effect of averaging learning curves over participants.</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>A typical ROC curve.</td>
<td>8</td>
</tr>
<tr>
<td>1.3</td>
<td>Equal variance signal detection model.</td>
<td>12</td>
</tr>
<tr>
<td>1.4</td>
<td>Unequal variance signal detection model.</td>
<td>14</td>
</tr>
<tr>
<td>1.5</td>
<td>The loss of parsimony in UVSD.</td>
<td>15</td>
</tr>
<tr>
<td>1.6</td>
<td>Dual-process signal detection model.</td>
<td>17</td>
</tr>
<tr>
<td>1.7</td>
<td>Averaging data from EVSD.</td>
<td>22</td>
</tr>
<tr>
<td>2.1</td>
<td>Hierarchical UVSD model</td>
<td>27</td>
</tr>
<tr>
<td>2.2</td>
<td>Effect of hierarchical structure.</td>
<td>29</td>
</tr>
<tr>
<td>2.3</td>
<td>Samples from marginal posterior distributions of the normal model.</td>
<td>37</td>
</tr>
<tr>
<td>2.4</td>
<td>Autocorrelation in criteria</td>
<td>42</td>
</tr>
<tr>
<td>2.5</td>
<td>Mitigation of autocorrelation</td>
<td>43</td>
</tr>
<tr>
<td>2.6</td>
<td>Relationships among the models.</td>
<td>49</td>
</tr>
<tr>
<td>2.7</td>
<td>Model analysis</td>
<td>53</td>
</tr>
<tr>
<td>2.8</td>
<td>Participant and item effects.</td>
<td>55</td>
</tr>
<tr>
<td>2.9</td>
<td>Word frequency effects</td>
<td>56</td>
</tr>
<tr>
<td>2.10</td>
<td>ROCs are invariant to transforms</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Parameter recovery of E-UVSD model</td>
<td>67</td>
</tr>
</tbody>
</table>
ABSTRACT

Recognition memory refers to a person’s ability to recognize something that has been previously encountered. For several decades recognition memory has been thought to be governed by a single process whereby the strength of a memory for an item dictates whether people judge the item as having been previously encountered or not. More recently, it has been proposed that recognition memory is governed by two, independent processes: Sometimes a memory judgement is based on strength, sometimes it is based on explicit recollection. Whereas this two-process theory has been embraced by many researchers, others claim that only one process is necessary to explain recognition memory. Here, I argue that all previous evidence for both the one and the two-process theories is questionable — because all models of recognition memory are non-linear models, averaging data over factors that vary (e.g., items) will distort the conclusions drawn. In all previous work it has been necessary to average data over items in order to fit formal models. To avoid the distortions from averaging, I develop hierarchical versions of popular recognition memory models that simultaneously account for person and item variability. These models are fit to data from several experiments to assess the veracity of previous claims. The results of this hierarchical modeling suggest that 1) ROC asymmetry, which has served as strong evidence for particular one and two-process model, is not an artifact of averaging, 2) The Yonelinas two-process model provides a superior account of recognition memory data when compared with the unequal-variance signal detection model via the DIC model-fit statistic, and 3) Two-process model fits reveal that estimates of recollection and familiarity co-vary across items and people. Moreover, manipulations of depth-of-processing, perceptual match-mismatch, response deadline, and list length all affect both recollection and familiarity to some degree. This result implies that, although the two-process model is the best-fitting parametric model, the data are
being generated from a yet-to-be specified one-process model.
Chapter 1

Introduction

The study of human memory has enjoyed a long and rich history in psychology. Recently, a large effort has been made to better understand the processes underlying recognition memory including representations of memory (Eich, 1982; Hintzman, 1984; Humphreys, Pike, Bain, & Tehan, 1989; Murdock, 1982), how recognition judgments are made (Glanzer & Adams, 1990; Morrell, Gaitan, & Wixted, 2002), and whether there are one or multiple memory processes (Wixted, 2007; Yonelinas & Parks, 2007). Testing whether there are one or multiple processes is becoming especially important, as dual process models are being utilized to understand deficits in memory due to various psychopathologies including Alzheimer’s disease (Bennett, Golob, Parker, & Starr, 2006), Parkinson’s disease (Davidson, Anaki, Saint-Cyr, Chow, & Moscovitch, 2006), schizophrenia (Breibion, David, Jones, & Pilowsky, 2005; Linscott & Knight, 2001), depression (Channon, Baker, & Robertson, 1993) and amnesia (Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998), and healthy aging (Dase-laar, Fleck, Dobbins, Madden, & Cabeza, 2006; Prull, Dawes, Martin, Rosenberg, & Light, 2006).

In science, it is critically important to understand how data may be used to test and constrain theory. In this dissertation, I argue that the link between theory and
data in the study of recognition memory processing is tenuous. The heart of this critique is that researchers must currently average recognition memory data over participants or items in order to evaluate competing theories. Often, these theories make fairly similar predictions about data patterns. I show that biases from averaging, even small ones, can provide spurious evidence for certain theoretical positions.

To avoid these misleading effects of averaging, models must explicitly account for variability from design elements such as the selection of participants and items. I construct several such models; some of these are classic recognition memory models that I extend to account for variability due to participants and the items that are studied, and some are novel theoretical developments. These models are then applied to new experiments in order to test benchmark results that have been based on fitting models to averaged data.

1.1 Averaging Data In Non-linear Models

Most models of cognitive processes are non-linear models — the mapping between the parameter space and the data is some nonlinear function. To demonstrate the effect of averaging data in such nonlinear models, let $\theta_1$ and $\theta_2$ be model parameters for two participants, let $f(\theta)$ be some nonlinear function that maps $\theta$ to the data space for both participants, and let $X_1$ and $X_2$ be the resulting observations for each participant. Note that the observations are generated from $f(\theta)$ without adding any noise; the only variance in the observations results from variance between participants.

Ideally, the data for each participant would be used to recover the parameters for each participant separately. If $f^{-1}(X)$ is the inverse function that maps data back into parameter space, then the separate estimates $\hat{\theta}_1 = f^{-1}(X_1)$ and $\hat{\theta}_2 = f^{-1}(X_2)$ are accurate. Moreover,

$$\frac{\theta_1 + \theta_2}{2} = \frac{f^{-1}(X_1) + f^{-1}(X_2)}{2},$$
that is, the average of the estimated parameters is the average of the true parameters.

In many cases, however, data are too sparse to estimate parameters for each participant separately, and so the data are averaged across participants and the averaged data are used to recover what is thought to be the averaged parameter values:

\[
\bar{X} = \frac{X_1 + X_2}{2},
\]

\[
\bar{\theta} = f^{-1}(\bar{X}),
\]

\[
\bar{\theta} \neq \frac{\theta_1 + \theta_2}{2}.
\]

As a concrete example, let the function that maps parameters to data be \(f(\theta) = \sqrt{\theta}\), and so the inverse function mapping the data back to parameters is \(f^{-1}(X) = X^2\).

Clearly, when data are not averaged parameters are recovered:

\[
X_1 = \sqrt{\theta_1},
\]

\[
X_2 = \sqrt{\theta_2},
\]

\[
\hat{\theta}_1 = X_1^2 = \left(\sqrt{\theta_1}\right)^2 = \theta_1,
\]

\[
\hat{\theta}_2 = X_2^2 = \left(\sqrt{\theta_2}\right)^2 = \theta_2.
\]

When the data are averaged, however, the parameter recovered from the averaged data is not the average of the true parameters:

\[
\bar{X} = \frac{\sqrt{\theta_1} + \sqrt{\theta_2}}{2},
\]

\[
\hat{\theta} = \bar{X}^2 = \left(\frac{\sqrt{\theta_1} + \sqrt{\theta_2}}{2}\right)^2,
\]

\[
\hat{\theta} \neq \frac{(\sqrt{\theta_1})^2 + (\sqrt{\theta_2})^2}{2}.
\]

(1.1)
Although this inequality is obvious, it is overlooked every time data are averaged across factors that vary in their parameters (e.g., participants or items). Moreover, in this example the distortions due to averaging are present even in the absence of any random noise included in the function mapping parameters to data. These distortions are, therefore, asymptotic: They will occur regardless of how much data are collected.

Because the function mapping parameters to data is non-linear, the magnitude and nature of the distortion resulting from averaging over participants depends on the values of the true parameters. In the above example, this simply means that the size of the difference between the left and right sides of the inequality in Equation 1.1 depends on the values of $\theta_1$ and $\theta_2$. In real analysis, true values are of course not known. The result is that the distortions due to averaging can not, for example, be estimated and accounted for with some correction. The only solution to avoiding them is to avoid averaging over things that vary, such as participants and items. Unfortunately, as the following examples show, doing so is not always easy.

1.1.1 An Example of Averaging in Learning

The consequences of averaging data can be seen clearly when learning curves are averaged across participants. The following seminal example is from Estes (1956). Consider an experiment in which participants are asked to repetitively perform a task, and learning is measured as decreases in reaction time across trials. Our goal is to determine whether learning is a discrete process (i.e., occurs all in one trial) or is continuous, occurring gradually across trials. Figure 1.1 shows hypothetical data in which each participant exhibits all-or-none learning, but the point at which learning occurs varies across participants. The learning curve constructed by averaging across participants (black line), however, clearly suggest that learning is gradual.

The effects of averaging data have also been explored within the context of process dissociation (Curran & Hintzman, 1995; Rouder, Lu, Morey, Sun, & Speckman,
Figure 1.1: The effect of averaging learning curves over people. Grey lines are reaction times for individuals across trials — each individual exhibits discrete learning but each learns at different rates. The black line is the averaged learning curve, which clearly suggests gradual learning.

2008) and two-choice signal detection (Rouder & Lu, 2005; Rouder et al., 2007). Before exploring the consequences of averaging recognition memory data, I present the recognition memory paradigm and competing models.

1.2 Recognition Memory Data

There are several variants of the recognition memory paradigm. Here I consider the tasks and resulting data that are most often used to compare competing recognition memory models.

1.2.1 Old-New Recognition

Recognition memory tasks usually consist of a study phase, a retention interval, and a test phase. During study, participants are shown a list of items and asked to remember them for a later test. Items are most often words. In some study tasks participants
may be instructed to simply read the words. In other tasks, participants may be asked to perform some task regarding the word, such as counting the number of vowels it contains. During study, participants are shown both words from the studied phase and new words that were not studied (one at a time and in a random order). The task is to decide whether each word is new or was studied. For each trial at study the item may be new or studied, and the participant may respond new or studied. Responses are coded as follows: a studied response to a studied item is a hit; a studied response to a new item is a false alarm. Hits and false alarms completely describe the data; their compliments (new responses to new and studied items) need not be considered.

Researchers commonly average responses over items to obtain hit and false alarm rates for each participant. Before considering how this averaging distorts results, it is useful to consider how these rates are commonly modeled. There are, at a minimum, two classes of attributes that determine the hit and false alarm rates. First, participants with high mnemonic abilities will have higher hit rates and/or lower false alarm rates than participants with poor mnemonic abilities. Second, some participants may be more biased than others toward a particular response, e.g., one may respond studied very often regardless of whether items were actually studied or not. The aim of all formal models is to characterize mnemonic ability and response bias; the differences between models is how they do so.

Each participant has a hit and false alarm rate that must be predicted by a model. All models, however, must include at least two parameters (one for mnemonic ability and one for bias). The result is that any model can perfectly predict the data, and so comparing the abilities of competing models to describe old-new recognition data makes little sense. Instead, researchers obtain more complex data, receiver operating characteristic curves, that may be used to distinguish models.
1.2.2 Receiver Operating Characteristic Curves

In order to provide data that can distinguish among models, each participant’s response bias can be manipulated systematically using differential payoffs (e.g., $10 per hit, $2 per correct rejection) or punishment. The resulting variability in hits and false alarm rates for each participant reflects differences in response biases with constant mnemonic ability.

These data are commonly presented as receiver operating characteristic (ROC) curves whereby the hit rate for each payoff condition is plotted as a function of that condition’s false alarm rate. The points in Figure 1.2 show a typically-observed ROC curve (with 5 payoff conditions). This curve shows two common properties of ROC curves. First, recognition memory ROCs are not straight lines, but rather exhibit substantial curvature. Second, ROCs are typically asymmetric about the negative diagonal, as if the curve were pushed up on the left and pushed down on the right (In Figure 1.2 the line shows a symmetric curve for comparison). Although this asymmetry is subtle, it is almost always observed (see Glanzer, Kim, Hilford, & Adams, 1999; Yonelinas & Parks, 2007, for reviews) and has driven an substantial body of theoretical development.

1.2.3 Confidence Ratings

Collecting ROC data by manipulating payoffs or punishments is costly as each participant must perform several trials at each level of payoff condition. Moreover, the usefulness of the data relies on participants being sufficiently biased by the payoff manipulation, which is notoriously difficult to accomplish\(^1\). An alternative way to construct ROC curves is by collecting participant’s confidence in their response that

\(^1\)We recently gave participants a 10-second ear-wrenching buzzing sound for every false alarm made, and no punishment for a miss. Whereas this punishment schedule should make any rational being avoid responding \textit{studied} unless they were absolutely sure an item was studied, the average false alarm rate was still 7%!
an item was new or studied. The most common design includes the six response options “sure new”, “believe new”, “guess new”, “guess studied”, “believe studied”, and “sure studied”. Theoretically, the differences between different levels of confidence for a participant reflect different levels of response bias in the same manner as manipulating payoff structure. Several researchers have constructed ROC curves from both payoff manipulations and confidence ratings, and find that they have the same shape (e.g., Egan, 1975; Ratcliff, Sheu, & Grondlund, 1992). This equivalence has recently been questioned (Broder & Schutz, 2009), however, Experiment 2 in Chapter 4 provides addition support for it. Most importantly, however, formal models of memory make clear and sometimes distinct predictions regarding the patterns of confidence-rating data, making them useful.

To demonstrate the construction of ROC curves from confidence rating data, let $K$ denote the number of response options, $S_k$ denote the proportion of studied-item trials for which the $k \in \{1, \ldots, K\}$ confidence judgement was made, and $N_k$ denote the same for new-item trials. For example, the proportion of “sure new” responses
to new items is $N_1$ and that of “sure studied” responses to studied items is $S_K$. The ratings to new and studied-item trials are transformed to false alarm rates $f_k$ and hit rates $h_k$, respectively, by cumulatively summing them from the “sure studied” rate to the “sure new” rate as follows:

\[ f_k = \sum_{l=K-k+1}^{K} N_l, \]  
\[ h_k = \sum_{l=K-k+1}^{K} S_l. \]

The values of $f_K$ and $h_K$ are always 1.0 and are omitted from ROC plots.

### 1.3 Models of Recognition Memory

In order to explore the effects of averaging recognition memory data, it is necessary to consider what models may be generating the data. In this section I provide a general overview of three popular models of recognition memory: 1) The equal-variance signal detection model, 2) The unequal-variance signal detection model, and 3) The Yonelinas dual-process model.

#### 1.3.1 Equal Variance Signal Detection

According to the theory of signal detection, all recognition judgements are made based on the strength of a particular memory trace. For example, when you are not sure which level you parked your car on, but looking at the “orange level” sign you have a faint feeling that it is on the “orange” level, your recognition is based on a weak memory. Likewise, if you see the “orange level” sign and are sure that you parked your car there, then you are experiencing a strong memory. This is a truly elegant theory of memory — it is termed a *single process model* as only one process, memory
strength, determines your recognition behavior (accuracy, speed, etc.).

On a recognition memory task, signal detection theory posits that each item at test elicits some amount of memory strength. On a given trial the response is determined by where the item’s strength lies relative to criteria set by the participant on the strength axis. The earliest signal detection model proposed to account for recognition memory data was the equal-variance signal detection (EVSD) model (Green & Swets, 1966; Kintsch, 1967). According to EVSD the latent mnemonic strengths of items are normally distributed with unit variance. The mean of the strength distribution is fixed to 0.0 for new items, and is a free parameter denoted $d'$ for studied items. In a two-choice task (i.e., respond “new” or “studied”) participants place a single criterion (denoted $C$) on the strength axis (see Figure 1.3A). If the strength of an item is above this criterion the participant responds “studied”; if the strength is below the criterion the participant responds “new”. According to the model, the probability of making “new” and “studied” responses to new items equals the area under the new-item strength distribution below and above the criterion, respectively. Likewise, the probability of making “new” and “studied” responses to studied items is the area below and above criterion under the studied-item distribution. These areas are defined by the standard normal cumulative distribution function (CDF) $\Phi(X)$:

\[
Pr(\text{False Alarm}) = Pr(\text{respond studied|new item}) = \Phi(-C)
\]
\[
Pr(\text{Hit}) = Pr(\text{respond studied|studied item}) = \Phi(d' - C)
\]

These equations make it clear that signal detection is a non-linear model — the mapping between parameters ($d'$ and $C$) and data (hits and false alarms) is the non-linear function $\Phi(X)$.

Within the EVSD framework parameter $d'$ measures a participant’s mnemonic
ability — larger distances between the new and studied-item distributions means it is easier to discriminate new from studied items based on their strength values. As $d'$ decreases and the distributions overlap, strength values are no longer diagnostic of whether an item was studied or is new. The criterion measures a participant’s response biases. For example, a person who is biased to response “studied” will have a low criterion, reflecting a high probability of both hits and false alarms.

The equal variance signal detection model may be easily extended to account for confidence ratings data. For $K$ response options, $K - 1$ criteria denoted $C_k$, $K = 1, \ldots, K - 1$, are placed on the strength axis (see Figure 1.3B). The probability of making a particular confidence response is simply the area under the normal distributions between the criteria. Let $X$ be confidence rating data where $X \in \{1, \ldots, K\}$ represents which response was made on a given trial. The mapping between EVSD parameters and confidence ratings data is as follows:

$$Pr(X = k|\text{new item}) = \Phi(C_k) - \Phi(C_{k-1})$$

$$Pr(X = k|\text{studied item}) = \Phi(d' - C_k) - \Phi(d' - C_{k-1})$$

where $C_0 = -\infty$, and $C_K = \infty$. For example, the probability that the second response (“believe new”) will be made on a new-item trial is the area under the standard normal distribution between the first and second criteria. For a studied-item trial the probability that the $K$th response (“sure studied”) will be made is the area under the normal distribution with mean $d'$ and variance 1.0 that lies above the highest ($K - 1$) criterion. The criteria in this model reflect several characteristics of participants’ responses biases. First, if a participant has a strong tendency to make the second confidence response (“believe new”), then the first and second criteria will be spread far apart. Second, an overall bias to respond “studied” is reflected by a concurrent increase in all criteria.
Figure 1.3: Equal variance signal detection model. A. Graphical representation of the EVSD model for a two-choice task. B. Graphical representation of the EVSD model for a confidence rating task with 6 response options. The new-item distribution is a standard normal, the studied-item distribution is normal with mean $d'$ and unit variance. C. ROC curves from EVSD model for $d'$ values ranging from .2 to 2.5.

ROC curves are plots of hits ($h_k$) as a function of false alarms ($f_k$). From the above equations, it can be easily shown that predicted ROCs for EVSD are:

$$h_k = \Phi(d' - \Phi^{-1}(f_k)).$$

where $\Phi^{-1}$ is the standard normal quantile function, and hit and false alarm rates are derived from confidence ratings per Equations 1.2 and 1.3. Predicted EVSD ROC curves for several values of $d'$ are shown in Figure 1.3C. These predicted ROC lines are curved in a manner consistent with empirical data. The ROCs predicted by EVSD, however, are symmetric about the negative diagonal. This failure of the EVSD model to account for ROC asymmetry has motivated two more complex models that can predict it. These models are extensions of EVSD, discussed in turn below.

### 1.3.2 Unequal-Variance Signal Detection

The first proposed extension of EVSD that can account for ROC asymmetry is to allow the variances of the new and studied-item distributions to differ (Egan, 1975). This model, termed the unequal-variance signal detection model (UVSD), is identical
to EVSD with the exception that the studied-item distribution has variance denoted $\sigma^2$ which is a free parameter (see Figure 1.4A). Predicted response probabilities are similar to those from EVSD:

$$
Pr(X = k | \text{new item}) = \Phi(C_k) - \Phi(C_{k-1})
$$

$$
Pr(X = k | \text{studied item}) = \Phi \left( \frac{d' - C_k}{\sigma} \right) - \Phi \left( \frac{d' - C_{k-1}}{\sigma} \right)
$$

From these equations the predicted UVSD ROC curve is:

$$
h_k = \Phi \left( \frac{d' - \Phi^{-1}(f_k)}{\sigma} \right)
$$

Predicted UVSD ROC curves for two values of $\sigma^2$ are shown in Figure 1.4B. As can be seen, the amount of asymmetry and $\sigma^2$ are positively related: the greater the variance in studied-item strength, the greater the ROC asymmetry. If $\sigma^2 = 1.0$ then UVSD reduces to EVSD, and thus predicts symmetric ROC curves. Figure 1.4C shows ROC data from a confidence ratings task, and the predicted ROC curve from EVSD and UVSD models. Although the difference (i.e., symmetry vs. asymmetry) is small, in almost every fit to data the value of $\sigma^2$ is found to be significantly greater than 1.0 (see Glanzer et al., 1999; Wixted, 2007; Yonelinas & Parks, 2007, for reviews).

The UVSD model has one more parameter, $\sigma^2$, than EVSD. Although this addition successfully predicts ROC asymmetry, it also comes with several undesirable consequences. The most important consequence is that the loss of parsimony over EVSD is great considering only one parameter has been added. This loss is even greater is ROC curves across conditions and people don’t share the same value of $\sigma^2$. In fact, there is substantial evidence that $\sigma^2$ varies across conditions (Glanzer et al., 1999). As a result, predicted ROC curves from UVSD vary in both $d'$ and $\sigma^2$. Figure 1.5 shows some such predictions for values of $d'$ and $\sigma^2$ that were generated randomly. Contrasting these with the predictions of EVSD (Figure 1.3C) clearly
shows the effect of adding only one extra parameter in UVSD. Nonetheless, the UVSD model is considered a single-process model as responses are determined solely by mnemonic strength.

### 1.3.3 Yonelinas’ Dual-Process Model

Single-process signal-detection models provided the foundation for recognition memory research for nearly 30 years. During this time, however, a growing body of evidence from other memory domains suggested that memory could only be described by two independent processes. For example, patients with extensive damage to the hippocampus exhibit an inability to form new memories (termed anterograde amnesia). One morning, Dr. Claparède (1911) unwittingly stabbed one of these patients with a needle during their routine morning handshake. The following morning, the patient did not remember ever meeting Dr. Claparède, however, the patient refused to shake his hand. The reason for the refusal was not that he been stabbed the previous morning, but rather, the patient made up fabricated reasons (e.g., fear of germs, hatred of doctors, etc.). These and similar findings lead researchers to conclude that
Figure 1.5: The loss of parsimony in UVSD. ROC curves were generated from the UVSD model with values of $d'$ and $\sigma^2$ were chosen from a Uniform(0,3).

Memory is divided into two separate systems: an explicit, conscious system, and an implicit, unconscious system.

Yonelinas (1994) proposed a two-process model account of recognition memory confidence-ratings data. According to this dual-process signal detection (DPSD) model, recognition judgments result from one of two independent processes. Sometimes, people have explicit, conscious recollection of an item. When this conscious recollection of a studied item occurs in a confidence-ratings task, the participant will always respond “sure studied”. Sometimes, however, conscious recollection may fail. When it does, responses are assumed to arise from implicit familiarity which is modeled as an equal-variance signal detection process. A graphical representation of this model is shown in Figure 1.6A. For studied items explicit recollection is an all-or-none process that occurs with probability $R$. If recollection fails, judgments are made based on implicit familiarity values arising from a normal distribution with mean $d'$ relative to criteria $C_k$. Judgments about new items are based solely on familiarity that arises from a standard normal distribution. The mapping between DPSD parameters and
confidence ratings data is as follows:

\[
\begin{align*}
\Pr(X = k \mid \text{new}) &= \Phi(C_k) - \Phi(C_{i(k-1)}) , \\
\Pr(X = k \in \{1, \ldots, K - 1\} \mid \text{studied}) &= (1 - R) \left[ \Phi(d' - C_k) - \Phi(d' - C_{(k-1)}) \right] , \\
\Pr(X = K \mid \text{studied}) &= R + (1 - R) \Phi(C_{(K-1)} - d') ,
\end{align*}
\]

The ROC curve predicted by the DPSD model is:

\[
h_k = R + (1 - R) \ast \Phi(d' - \Phi^{-1}(f_k)).
\]

The probability of recollection serves as a y-intercept for the ROC curve, and the remainder of the curve is traced out as it is in the equal-variance signal detection model. Figure 1.6B shows predicted ROC curves for two levels of recollection. The amount of explicit recollection is positively related to ROC asymmetry: the more recollection, the more asymmetry. When recollection is absent \((R = 0)\) DPSD reduces to an equal-variance signal detection model. Thus, values of \(R > 0\) index the extent to which the ROC curve deviates from symmetry. This relationship is similar to that between asymmetry and \(\sigma^2\) in the UVSD model. The difference is that the dual-process model posits that the amount of ROC asymmetry provides a direct measure of explicit recollection.

The dual-process model of recognition memory has been used to examine deficits in memory, with claims often that a deficit affects one or the other process, or both. Nevertheless, many theorists conclude that recognition memory is mediated by a single process, and that these researchers are incorrectly interpreting ROC asymmetries (Dunn, 2008; Glanzer et al., 1999; Heathcote, Raymond, & Dunn, 2006; Wixted, 2007, e.g.). In the following sections, I present a novel argument for why the single-process view of memory is more plausible than previously thought.
1.4 Separating Process and Nuisance Variability

Recognition responses (e.g., which confidence rating is made) vary from trial-to-trial. This variability is in part due to variability in process (e.g., recollection occurring or not). Trial-to-trial variability, however, also arises because each trial is a unique combination of a participant and an item. Each participant has a unique mnemonic ability and response bias. Items also differ in sensitivity and bias. For example, an item’s word frequency (how often the word appears in common texts) has a large effect on its ability to be remembered (Peters, 1936; Scarborough, Cortese, & Scarborough, 1977; Schwartz & Rouse, 1961; Shepard, 1967). It is critical to separate variability due to the mnemonic process from that due to items and participants. Many researchers, however, have overlooked this point, especially on what variability in process entails.

To understand variability in process, consider the hypothetical situation in which a single participant is tested on a single item. If we could perform this test over and over, without any learning or testing effects, the variability in response would reflect the mnemonic process for this participant and item combination. The underlying patterns in this variability, such as whether the resulting confidence rating ROCs are
curved or straight lines, or are asymmetric, reflect deep structural properties that are not a function of the person or item. An example of such a property is that recognition memory is mediated by a mixture of recollection and familiarity. In fact, the elucidation of these structural properties, uncontaminated by variability in people or items, is the main object of study. Separate from these deep structural properties is variation from items and participants. For instance, the probability of recollection in DPSD may vary across items.

The ramifications of this fact are subtle, especially when discussing the UVSD model. Mickes, Wixted, and Wais (2007) and Wixted (2007), for example, speculate that the asymmetry in ROC plots reflects item variability rather than a deep structural property. They note that items will differ in how much strength is gained at study, and when data are averaged, this difference will result in greater variability across studied items than new. This scenario, however, is not the only possibility. It may be that the asymmetry is a deep structural property that does not reflect item variability but is instead a signature of process variability. In fact, the dual-process model is a process-variability explanation because recollection is assumed to occur and add asymmetry even in the absence of item variation. In the next section I provide a concrete example of how averaging data over items confounds item variability with variability in process.

1.5 The Effects of Averaging In Recognition Memory

The effects of averaging data in any non-linear context depend on what the true generating model is, the true parameters of that model, and the extent to which those parameters vary. Obviously we are not privy to the true model or its parameters, but we can consider some hypothetical cases. For example, it is well known that
words vary in their memorability (e.g., due to word-frequency differences). In signal
detection, these effects would be manifest as different values of $d'$ across items. Rouder
and Lu (2005) and Rouder et al. (2007) showed that in old-new recognition, such
variability in $d'$ will bias estimates of $d'$ to be too small, and bias confidence intervals
on those estimates to be too small. Here I demonstrate the effects of averaging
confidence-ratings data with item variability, first considered in Morey, Pratte, and

Imagine that recognition memory is mediated by the simple equal-variance signal
detection model presented previously. Consider also that the stimuli consist of easy
and hard items (low and high $d'$ values, respectively). In order to construct ROC
curves and estimate model parameters the data are averaged over the easy and hard
items. Fitting the UVSD and DPSD models to these averaged data should provide
estimates of $\sigma^2 = 1.0$ and $R = 0$, respectively.

To make the situation concrete, let $d'$ for the easy and hard items be 3 and 1,
respectively. For a criterion of 1.0 data can be generated from the true model with
averaged parameter $d' = 2$ as follows:

\[
F_{true} = \Phi(-1) = .16
\]

\[
H_{true} = \Phi(2 - 1) = .84
\]

In real analysis, data are not generated from the average model, but rather, data are
generated differently for the easy and hard items, and these data are averaged:

\[
F_{\text{hard}} = \Phi(-1) = .16 \\
H_{\text{hard}} = \Phi(1 - 1) = .50 \\
F_{\text{easy}} = \Phi(-1) = .16 \\
H_{\text{easy}} = \Phi(3 - 1) = .98 \\
F_{\text{averaged}} = \frac{.16 + .16}{2} = .16 \\
H_{\text{averaged}} = \frac{.50 + .98}{2} = .74
\]

The false alarm rates are the same for the true (average) model and the averaged data because different values of \(d'\) do not affect false alarm rates. The hit rate obtained by averaging data, however, is substantially lower that that generated from the true (average) model. The result is that performance will be underestimated; this result holds across a range of true \(d'\) values and levels of variability (Rouder et al., 2007). Critically, note that in this demonstration no noise other than that between items was introduced. The fact that the bias in sensitivity is present even in the absence of random noise shows that it is asymptotic: the results from averaged data will be wrong regardless of how much data are collected.

The above demonstration can be extended to ROC data by completing the same data-generation and averaging procedures for a range of criteria. The results of this demonstration are displayed in Figure 1.7. The solid line is the ROC generated from the true model; EVSD with \(d' = 2.0\). The dashed line shows the ROC generated by averaging hit and false alarm rates over easy and hard items. Clearly this data-averaged ROC does not reflect the averaged generating model. The most alarming aspect of this data-averaged ROC is that it is asymmetric about the negative diagonal in exactly the manner predicted by UVSD and DPSD. In fact, fitting the UVSD model to averaged data reveals an estimate of \(\sigma^2 = 1.72\); similar to what is commonly
observed. Likewise, fitting the dual-process model to these data reveals an estimate of recollection $R = .30$, a value also in line with common findings. Again, this bias resulting from averaging over items that vary is asymptotic: it will occur regardless of how much data is collected, and replicating results (e.g., asymmetry) will only replicate the bias. The result is that, when data are averaged over items that vary, variability in those items is confounded with variability in process, such that item variability will be mistaken for evidence for processes such as recollection.

The distortions due to averaging have been acknowledged by a minority of authors including Malmberg and Xu (2006), Ratcliff, McKoon, and Tindall (1994), and Wixted (2007). Nevertheless, in all previous studies of recognition memory, data were averaged over people, items, or both in order to construct ROC curves and/or fit formal models. As a result, all support for models that are more complex than EVSD should be viewed as tenuous at best. These models include both UVSD and DPSD.

The result of averaging in Figure 1.7 is only one of an infinite number of possibilities. For example, perhaps there is more or less variability in $d'$ than in the example; perhaps there is variability in response biases (criterion). Alternatively, perhaps the dual-process model is the true generating model. If so, then ROC asymmetry is real, but data averaged from the dual-process model will be distorted in other ways. Figure 1.7 shows a worst-case scenario: a simple model being rejected for complex models due to bias, and that bias being interpreted as a meaningful characteristic of the mnemonic system. Regardless of what the true generating process is, however, data averaged over participants or items will be distorted in some way.

Constructing ROCs and fitting formal models using conventional techniques requires that data are averaged over something (people or items) in order to construct hit and false alarm rates. In order to avoid distortions from averaging, it is necessary to specify models at the participant by item level. For example, in EVSD each person
and each item must be free to have different $d'$ and criteria parameters. Doing so is far from trivial, and is only plausible due to very recent advances in statistics and computer power. These difficulties reveal why previous researchers have been content to average data: the alternative is hard.

In the following chapters I develop EVSD, UVSD, DPSD, and other recognition memory models that explicitly account for participant and item variation. These models provide the first accurate picture of what recognition memory data look like without distortions from averaging. In Chapter 2 I develop the UVSD model and use it to determine whether ROC curves are truly asymmetric (i.e., $\sigma^2 > 1.0$) when participants and items are not averaged over as in Figure 1.7. The results reveal that ROC asymmetry is real. In Chapter 3 I develop a dual-process model that accounts for participant and item variability, and compare it to the UVSD model. The results reveal that the dual-process model provides a better fit to the data than the UVSD model. The results also suggest, however, that a single-process model other than UVSD may
be able to account for the data. Given the success of the fit of the dual-process model in Chapter 3, in Chapter 4 I explore whether the recollection and familiarity parameters of the dual-process model are truly independent, by assessing whether they can be selectively influenced by different manipulations. The results indicate that, when participant and item variability are properly accounted for, manipulations of study list length reduce estimates of recollection, whereas levels of familiarity are relatively preserved. In addition, a levels-of-processing manipulation is shown to greatly affect recollection estimates but affect familiarity only slightly. The results suggest that, perhaps, two processes are needed to account for recognition memory, and that these may be recollection and familiarity as defined in the Yonelinas dual-process mode.
Chapter 2

Accounting for Participant and Item Effects in UVSD

In Chapter 1 it was shown that averaging data over participants or items may result in ROC asymmetry even when the data were generated from an equal-variance signal detection model. To avoid this bias it is necessary to measure ROC asymmetry without confounding it with person or item variability. Because the value of $\sigma^2$ in UVSD is a measure of ROC asymmetry, an estimate of $\sigma^2$ from a UVSD model that accounts for participant and item variability provides an unbiased measure of ROC asymmetry. If ROC curves are symmetric, then this uncontaminated estimate of $\sigma^2$ will equal 1.0. Alternatively, if this estimate of $\sigma^2$ is greater than 1.0, then asymmetry is a feature of the mnemonic process and not a result of item variability.

In this Chapter I develop the hierarchical UVSD model, which explicitly accounts for participant and item variability. This work is based on Morey, Pratte, and Rouder (2008) and Pratte et al. (in press). This model is similar to the signal detection model for two-alternative forced-choice data developed in Rouder and Lu (2005) and Rouder et al. (2007), but can account for confidence ratings data.

The UVSD model is used here as a psychometric model for measuring ROC asym-
metry. The key parameter is $\sigma^2$; if ROCs are symmetric, then $\sigma^2 = 1.0$. As asymmetry in ROCs increases, $\sigma^2$ increases above 1.0. The use of UVSD is motivated by a number of factors: First, UVSD was the first, and remains the most straightforward generalization of signal detection theory made to account for asymmetry. Second, UVSD has been shown repeatedly to provide a good fit to averaged data (e.g., Glanzer et al., 1999; Heathcote, 2003; Slotnick & Dodson, 2005; Wixted, 2007) and thus provides a logical starting point for fitting non-averaged data. Third, providing a hierarchical extension of UVSD is both feasible and convenient, as evidenced by the two-alternative model developed in Rouder et al. (2007). In theory, it is possible that other models, such as Yonelinas’ dual process model, may be extended hierarchically to measure asymmetry. In DPSD the estimate of recollection serves as the index of asymmetry. This model is developed in Chapter 3.

2.1 Hierarchial UVSD Model Specification

Consider a recognition memory test in which each of $I$ participants ($i = 1, \ldots, I$) is tested on each of $J$ items ($j = 1, \ldots, J$). According to signal detection theory each item at test gives rise to some latent mnemonic strength denoted $X_{ij}$. Participants are assumed to place criteria on the latent strength dimension such that the probability of making the $k$th response is equal to the probability that the latent strength falls between the $k-1$ and $k$th criterion. Typically, the mean of the new-item distribution is set to 0.0 to fix the location of the strength space. It is more convenient in accounting for item and participant effects to allow this mean to be a free parameter and center the space on a criterion (see Rouder et al., 2007). A graphical representation of the UVSD model with this parameterization is shown in Figure 2.1A. Latent strengths
are given by:

\[ X_{ij} \sim \begin{cases} 
\text{Normal}(d_{ij}^{(n)}, 1) & \text{new,} \\
\text{Normal}(d_{ij}^{(s)}, \sigma^2) & \text{studied,}
\end{cases} \]

where \( d_{ij}^{(s)} \) and \( d_{ij}^{(n)} \) are means for the \( ij \) person-by-item combination when the item is studied and new, respectively, and \( \sigma^2 \) is the studied-item variance that is assumed to be the same across all participants and items (this assumption is tested in Chapter 3).

Criteria \( c_{ik} \) are assumed to vary across participants, but not across items. This variability in individual criteria reflects participants’ biases for certain responses over others. The probability that the \( i \)th person makes the \( k \)th response to the \( j \)th item is:

\[
\Pr(y_{ij} = k \mid \text{new}) = \Phi\left(d_{ij}^{(n)} - c_{ik}\right) - \Phi\left(d_{ij}^{(n)} - c_{i(k-1)}\right),
\]

\[
\Pr(y_{ij} = k \mid \text{studied}) = \Phi\left(\frac{d_{ij}^{(s)} - c_{ik}}{\sigma^2}\right) - \Phi\left(\frac{d_{ij}^{(s)} - c_{i(k-1)}}{\sigma^2}\right),
\]

where \( c_0 = -\infty, c_{K/2} = 0 \) (\( C_{(K+1)/2} = 0 \) if \( K \) is odd), \( c_K = \infty \), and \( \Phi \) is the CDF of the standard normal.

In this parametrization sensitivity \( d' \) is the distance between the new and studied-item distributions \( d'_{ij} = d_{ij}^{(s)} - d_{ij}^{(n)} \). Overall response biases (e.g., to response “studied”) are concurrent shifts in the distributions (see Figure 2.1B). It has been shown that differences between conditions often manifest as concurrent increases in hit rates and decreases in false alarm rates (Glanzer & Adams, 1990; Glanzer, Adams, Iverson, & Kim, 1993; Stretch & Wixted, 1998). This pattern, termed the mirror effect, suggests that in conditions in which items are more easily recognized as studied when they were studied, items that are new are also better recognized as such. In the parameterization used here, mirror effect are negative correlations between the new
and studied-item distributions (see Figure 2.1C).

Figure 2.1: Hierarchical UVSD model. A. Hierarchical unequal-variance signal detection model with middle criteria fixed to 0.0 and the mean of the new-item distribution \((d^{(n)})\) free to vary. B. Response biases are represented as concurrent shifts, or positive correlations, in the new and studied-item means. C. Mirror effects present as negative correlations in the new and studied-item means.

The means of the new and studied-item distributions can not be estimated without some restriction, as each participant-by-item combination occurs only once, and only as either new or studied. To make the model estimable, additive structures are placed on the means:

\[
d^{(n)}_{ij} = \mu^{(n)} + \alpha^{(n)}_i + \beta^{(n)}_j, \tag{2.1}
\]

\[
d^{(s)}_{ij} = \mu^{(s)} + \alpha^{(s)}_i + \beta^{(s)}_j + \theta^{(s)} l_{ij}, \tag{2.2}
\]

where \(\mu^{(n)}\) and \(\mu^{(s)}\) are grand means, \(\alpha^{(n)}_i\) and \(\alpha^{(s)}_i\) are participant effects, and \(\beta^{(n)}_j\) and \(\beta^{(s)}_j\) are item effects. In addition to participant and item effects, recognition responses may be effected by the number of items that intervene between the study and test of an item, termed the lag for that item. Parameter \(\theta^{(s)}\) is the linear effect of this study-test lag \(l_{ij}\).

Participant and item effects are treated as random effects — they are considered samples from parent distributions. The following parent distributions are placed on
participant and item effects:

\[ \alpha_i^{(n)} \sim \text{Normal}(0, \sigma_{\alpha,n}^2), \]  
\[ \beta_j^{(n)} \sim \text{Normal}(0, \sigma_{\beta,n}^2), \]  
\[ \alpha_i^{(s)} \sim \text{Normal}(0, \sigma_{\alpha,s}^2), \]  
\[ \beta_j^{(s)} \sim \text{Normal}(0, \sigma_{\beta,s}^2). \]  

The variances \( (\sigma_{\alpha,n}^2, \sigma_{\beta,n}^2, \sigma_{\alpha,s}^2, \sigma_{\beta,s}^2) \) are free parameters; their magnitudes reflect the size of participant and item variability.

The effect of the hierarchical structures is to shrink effect estimates toward zero. This shrinkage will be largest for extreme effects—the interpretation is that extreme effects are the result of noise, rather than having been generated from extreme true values. The amount of shrinkage depends on how much evidence there is for extreme effects, determined both by how much data there are and how well those data inform parameter estimates. For example, Figure 2.2 shows estimates of sensitivity \( d' \) for each of the 480 items in Experiment 1 (presented below) from the hierarchical model, as a function of estimates obtained from an equivalent non-hierarchical fixed-effects model. Shrinkage is seen as pulling items with small \( d' \) estimates up (toward the average item), and pulling items with large estimates down. Although both the hierarchical and non-hierarchical versions have the same number of parameters (one for each item), the hierarchical model is more constrained as parameters are not as free to vary from one another. Random-effects modeling has been preferred for hierarchical modeling of cognitive phenomena because it provides a more parsimonious model as well as a means of generalizing results to a larger class of participants and items (e.g., Lee, 2006; Morey, Rouder, & Speckman, 2008, 2009; Rouder et al., 2007; Rouder, Morey, et al., 2008; Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008). In chapter 3 the choice to model effects hierarchically by letting effect variances be a free parameter
Hierarchical d’ presented in Experiment 1. Item sensitivities are defined as $\mu^{(s)} - \mu^{(n)} + \beta_j^{(s)} - \beta_j^{(n)}$. Estimates on the y-axis are from the hierarchical model presented in Chapter 1. Estimates on the x-axis are from the same model, but with no hierarchical pooling. This non-hierarchical model was achieved by setting the priors on effect variances $\sigma_{\beta(n)}^2$ and $\sigma_{\beta(s)}^2$ to 2.0, as opposed to the hierarchical model in which these parameters were estimated as .20 and .16, respectively.

is compared with non-hierarchical versions in which variance parameters are fixed to relatively large values.

### 2.2 Model Analysis

The hierarchical UVSD model is conveniently analyzed in the Bayesian framework with Gibbs sampling (Gelfand & Smith, 1990). In this section, I provide a brief introduction to Bayesian analysis, justify priors for the UVSD model, present the full conditional posterior distributions of each parameter, and provide sampling algorithms for estimating the marginal posterior distributions of the parameters. I have developed the R library **hbmem** which performs the model analysis in a fully auto-
mated fashion. This library may be installed by issuing `install.packages(hbmem)` at the R prompt. The included function `simUVSD` will simulate data from the hierarchical UVSD model developed above, the function `sampleUVSD` will fit the hierarchical UVSD model to those data using the procedures outlined below.

### 2.2.1 An Introduction to Bayesian Analysis

In order to briefly introduce Bayesian analysis I develop a simple normal model (see Rouder and Lu (2005) or Gelman, Carlin, Stern, and Rubin (2004) for more in-depth treatments). Consider the case in which \( X_i, \ i = 1, \ldots, I \), are normally distributed with unknown mean \( \mu \) and variance \( \sigma^2 \):

\[
X_i \sim \text{Normal}(\mu, \sigma^2)
\]

The goal of Bayesian analysis is to determine the distribution of the unknown parameters \( \mu \) and \( \sigma^2 \) given the data \( X_i \). This quantity is termed the *posterior distribution* of the parameters. The likelihood function for the model is the probability of the data, given the parameters. Bayes’ theorem gives the relationship between these two different conditional statements:

\[
f(\mu, \sigma^2 | X_i) = \frac{f(X_i | \mu, \sigma^2) f(\mu, \sigma^2)}{f(X_i)}
\]

The term \( f(\mu, \sigma^2) \) is the *prior distribution* of the parameters, and specifies our prior beliefs about what values the parameters may be, before observing any data. The goal is to find the term on the left, the joint posterior distribution of the parameters. To do so we only need to consider terms on the right that involve the parameters of interest (everything else cancels out in integration). The posterior then need only be
defined up to constants of proportionality with respect to the parameters:

\[ f(\mu, \sigma^2|X_i) \propto f(X_i|\mu, \sigma^2) f(\mu, \sigma^2) \]  

(2.7)

This equation states that the distribution of the parameters given the data is proportional to the likelihood \( f(X_i|\mu, \sigma^2) \) times the prior \( f(\mu, \sigma^2) \). For the normal model the likelihood is:

\[ f(X_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( \frac{\sum_{i=1}^t (X_i - \mu)^2}{-2\sigma^2} \right) \]

The prior distribution \( f(\mu, \sigma^2) \) is chosen before analysis. There are many options for priors in this example, and there are good reasons for choosing some over others. Here I choose the distributional form of priors based on how convenient they are, and choose the parameters of those forms so that priors are relatively uninformative, that is, they have little influence on estimates. Parameters \( \mu \) and \( \sigma^2 \) are independent, and so the joint prior \( f(\mu, \sigma^2) \) is simply the product of the marginal priors on \( \mu \) and \( \sigma^2 \) which are:

\[ \mu \sim \text{Normal}(\mu_0, \sigma_0^2) \]

\[ \sigma^2 \sim \text{IG}(\alpha, \beta) \]

where IG is the inverse gamma distribution with scale \( \alpha \) and shape \( \beta \). Prior parameters \( \mu_0, \sigma_0^2, \alpha, \) and \( \beta \) are chosen before analysis to reflect how much knowledge we have about each parameter. To make the prior on \( \mu \) weakly informative the value of \( \sigma_0^2 \) is set to a large number. To make the prior on \( \sigma^2 \) weakly informative the values of both \( \alpha \) and \( \beta \) are chosen to be near zero. According to this specification the form
of the priors are:

\[ f(\mu | \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left( \frac{(\mu - \mu_0)^2}{-2\sigma_0^2} \right) \]

\[ f(\sigma^2 | \alpha, \beta) = \frac{\alpha^\beta}{\Gamma(\beta)} \left( \frac{1}{\sigma^2} \right)^{\beta+1} \exp \left( -\frac{\alpha}{\sigma^2} \right) \]

The joint prior \( f(\mu, \sigma^2) \) is simply the product of these marginal priors. Multiplying the likelihood and the priors according to Equation 2.7 yields the joint posterior. Again, in constructing the posterior we can ignore all terms that are multiplicative constants with respect to the parameters, yielding:

\[ f(\mu, \sigma^2 | X_i, \mu_0, \sigma_0^2, \alpha, \beta) \propto \exp \left( \sum_{i=1}^{I} (X_i - \mu)^2 \right) \exp \left( \frac{(\mu - \mu_0)^2}{-2\sigma_0^2} \right) \left( \frac{1}{\sigma^2} \right)^{\beta+1} \exp \left( -\frac{\alpha}{\sigma^2} \right) \]

Equation 2.8 is the joint posterior distribution of parameters \( \mu \) and \( \sigma^2 \) given some data \( X_i \) and prior parameters. Our main interest is not in the full joint posterior, but rather, we would like to know the marginal distributions of \( \mu \) and \( \sigma^2 \). In fact, parameters like \( \sigma^2 \) are often of no interest at all, and our wish is to learn about parameter \( \mu \) without regard to or interest in \( \sigma^2 \). Finding the marginal posteriors of parameters from the joint is a matter of integrating the joint posterior over all parameters except the one of interest:

\[ f(\mu) \propto \int_0^\infty f(\mu, \sigma^2 | X_i, \mu_0, \sigma_0^2, \alpha, \beta) \, d\sigma^2. \]

\[ f(\sigma^2) \propto \int_{-\infty}^\infty f(\mu, \sigma^2 | X_i, \mu_0, \sigma_0^2, \alpha, \beta) \, d\mu. \]

In most applications, including this one, it is not possible to evaluate these integrals analytically. Instead, they may be evaluated computationally using Markov chain Monte Carlo (MCMC) integration. In MCMC integration samples are drawn from the conditional posterior distribution of each parameter, i.e., the parameter conditioned
on a random draw from the posterior of all other parameters. Conditional posteriors are easily derived from the joint posterior. For example, the conditional posterior of $\mu$ given $\sigma^2$ is the terms of the full joint posterior involving $\mu$:

$$f(\mu|X_i, \sigma^2, \mu_0, \sigma_0^2) \propto \exp \left( \frac{\sum_{i=1}^{I} (X_i - \mu)^2}{-2\sigma^2} \right) \exp \left( \frac{(\mu - \mu_0)^2}{-2\sigma_0^2} \right)$$

which here is the likelihood and the prior on $\mu$. As stated, in MCMC the goal is to draw samples of $\mu$ from this conditional distribution. There are many techniques for sampling from unknown distributions such as the Metropolis-Hastings algorithm. In this case, though, the conditional posterior of $\mu$ is actually a normal distribution. Expanding the polynomials and collecting terms, the posterior can be rewritten as:

$$f(\mu|X_i, \sigma^2, \mu_0, \sigma_0^2) \propto \exp \left( -\frac{1}{2} \left[ \mu^2 \left( \frac{I}{\sigma^2} + \frac{1}{\sigma_0^2} \right) - 2\mu \left( \frac{IX}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \right] \right).$$

Let $\lambda = \left( \frac{IX}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)$ and $\tau = \left( \frac{I}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1}$. Completing the square yields:

$$f(\mu|X_i, \sigma^2, \mu_0, \sigma_0^2) \propto \exp \left( \frac{(\mu - \lambda\tau)^2}{-2\tau} \right)$$

which is (up to a constant of proportionality with respect to $\mu$) a normal distribution:

$$f(\mu|X_i, \sigma^2, \mu_0, \sigma_0^2) \sim \text{Normal}(\lambda\tau, \tau) \quad (2.9)$$

In this example, if $\sigma_0^2$ is set to $\infty$, it is clear that the posterior distribution of $\mu$ is the sampling distribution of $\mu$ in traditional frequentist statistics. Choosing large values of $\sigma_0^2$ thus approximates the frequentist treatment of $\mu$.

The conditional posterior for $\sigma^2$ can be shown to be an inverse gamma distribution
with a similar exposition as for \( \mu \) above:

\[
f(\sigma^2|X_i, \mu, \alpha, \beta) \sim IG \left( \frac{I}{2} + \alpha, \frac{\sum_{i=1}^{I}(X_i - \mu)^2}{2} + \beta \right)
\] (2.10)

Having derived the conditional distributions of each parameter, we can now use MCMC to estimate the marginal distributions. To do so MCMC includes the following steps:

**Step 1.** Choose starting values for \( \mu \) and \( \sigma^2 \). Although choosing starting values that are close to true values will speed estimation, in most situations the choice is arbitrary.

**Step 2.** Sample \( \mu \) from its conditional distribution in Equation 2.9, with \( \sigma^2 \) set to its starting value.

**Step 3.** Sample \( \sigma^2 \) from its conditional distribution in Equation 2.10 with the value of \( \mu \) set to that just sampled in step 2.

**Step 4.** Take another sample of \( \mu \), but now with the value of \( \sigma^2 \) set to that just sampled in step 4.

**Step 5.** Repeat steps 3 and 4 several of times. The number of times are referred to as the *iterations* of the MCMC loop.

These steps may be easily implemented in any programming language that is capable of drawing random samples from normal and gamma distributions. I provide R code for demonstration:
# Make Data
I=50 # Number of samples
X=rnorm(I,100,sqrt(15)) # Generate Data
xbar=mean(X)

# Set Priors
mu0=0
sig20=100
a=b=.01

# Prepare For MCMC Chain
M=1000 # Number of MCMC iterations
s.mu=rep(NA,M) # Save space for samples of mu
s.sig2=rep(NA,M) # Save space for samples of sigma2
s.mu[1]=90 # Starting value for mu
s.sig2[1]=5 # Starting value for sigma2

for(m in 2:M) # Begin MCMC Loop
{
# Sample mu, given last s.sig2
lambda=I*xbar/s.sig2[m-1] + mu0/sig20
tau=(I/s.sig2[m-1] + 1/sig20)^(-1)
s.mu[m]=rnorm(1,lambda*tau,sqrt(tau))

# Sample sigma2, given just sampled s.mu
SS=sum((X-s.mu[m])^2)
s.sig2[m]=1/rgamma(1,shape=I/2 + a,rate=SS/2 + b)
}
The result of the MCMC loop are samples from the marginal posterior distributions of $\mu$ and $\sigma^2$. Figures 2.3A and B show these samples for $\mu$ and $\sigma^2$, respectively, for each iteration of the chain. As can be seen, the chains get very near their true values quickly. This is referred to as convergence, and shows why the choice of starting value is unimportant in this example. After looking at the chain in this manner, several of the first samples are discarded until it is clear that the chain has converged. This burnin period ensures that all samples are actually from the posterior, and are not influenced by our choice of the starting values. Means of these samples, termed posterior means, serve as point estimates of parameters. The variance of the samples characterizes our uncertainty in the parameter estimates. Similarly, quantiles of the samples serve as credible intervals. For example, the .025 and .975 quantiles describe the region that contains 95% of the posterior mass. For $\mu$, the posterior mean for these data is 99.45; the 95% credible interval is (98.32, 100.56). Given these data, we are 95% confident that the true value of $\mu$ lies within this region. Rather than rely on summary statistics, Figures 2.3C and D show smoothed histograms derived from the sampled marginal distributions. These are the (approximated) marginal distributions of $\mu$ and $\sigma^2$ given the data.

The steps for estimating more complex models follow the same procedures outlined above. Sometimes, however, conditional posteriors do not follow a known form such as the Normal, and must be sampled with more specialized techniques. Moreover, often the chains do not converge so rapidly as those in the simple normal model, and so many MCMC iterations with a large burnin is needed. Finally, for many models there is autocorrelation in the samples of a parameter, such that each draw is not an independent draw from the posterior. When this happens steps must be taken to mitigate the autocorrelation. In all cases, however, the end result is samples from the marginal posterior distribution of each parameter.
Figure 2.3: Samples from marginal posterior distributions of the normal model. **A.** Chain from samples of $\mu$ across 1000 MCMC iterations. **B.** Chain from samples of $\sigma^2$ across 1000 MCMC iterations. **C.** Smoothed histogram of samples of $\mu$. **D.** Smoothed histogram of samples of $\sigma^2$.

### 2.2.2 Priors for UVSD Model

Normal priors are placed on the grand means of the new- and studied-item distributions and lag effects:

\[
\mu^{(n)}, \mu^{(s)}, \theta^{(s)} \overset{iid}{\sim} \text{Normal}(0, \sigma^2). 
\]
Inverse gamma priors are placed on all variance parameters:

\[
\sigma^2 \sim \text{IG}(a, b),
\]

\[
\sigma_{\alpha(n)}^2, \sigma_{\beta(n)}^2, \sigma_{\alpha(s)}^2, \sigma_{\beta(s)}^2 \sim \text{IG}(e, f).
\]

Values of \(\sigma_\mu^2\), \(a\), \(b\), \(e\) and \(f\) must be chosen before analysis. As long as the value of \(\sigma_\mu^2\) is sufficiently large, the prior is approximately non-informative. I use \(\sigma_\mu^2 = 100\) though other large values will yield the same results. Values of \(a = e = 2\) and \(b = f = 1\) work well. With these choices for \((a, b, e, f)\), the priors are vaguely informative, but this information has a minimal effect on the posterior means for variance parameters. Simulation studies reveal that, given a reasonable sample size, the model converges with less informative priors (e.g., \(\sigma_\mu^2 = 1000\), \(a = b = e = f = .01\)), and more informative priors (e.g., \(\sigma_\mu^2 = 10\)) have only negligible effects on estimates. For smaller sample sizes (e.g., 25 participants), the choice of priors becomes more important, and the effect of various choices should be explored through simulation.

### 2.2.3 Conditional Posterior Distributions

The following notation is used in deriving the conditional posterior distributions. Parameters in bold type indicate a vector of parameters or data; for example, \(y\) denotes the vector of all data. Let \(s_{ij}\) indicate whether the \(j\)th item for the \(i\)th participant was studied (\(s_{ij} = 1\)) or not (\(s_{ij} = 0\)). Derivation of conditional posterior distributions is greatly aided by introducing a set of latent variables. Following Albert and Chib (1993), let \(w_{ij}\) be related to \(y_{ij}\) as follows:

\[(y_{ij} = k) \iff (c_{i(k-1)} \leq w_{ij} < c_{ik}).\]
Random variables $w_{ij}$ are distributed as normals:

$$w_{ij} \overset{\text{indep.}}{\sim} \begin{cases} \text{Normal}(d^{(n)}_{ij}, 1), & s_{ij} = 0, \\ \text{Normal}(d^{(s)}_{ij}, \sigma^2), & s_{ij} = 1. \end{cases}$$

With these definitions, it is obvious that

$$Pr(y_{ij} = k) = Pr(c_{i(k-1)} \leq w_{ij} < c_i k) = \begin{cases} \Phi\left(\frac{d^{(n)}_{ij} - c_{i(k-1)}}{c_i(k-1)}\right) - \Phi\left(\frac{d^{(n)}_{ij} - c_{i(k-1)}}{c_i k}\right), & s_{ij} = 0, \\ \Phi\left(\frac{d^{(s)}_{ij} - c_{i(k-1)}}{\sqrt{\sigma^2}}\right) - \Phi\left(\frac{d^{(s)}_{ij} - c_{i(k-1)}}{\sqrt{\sigma^2}}\right), & s_{ij} = 1, \end{cases}$$

The latent data are introduced because it is more convenient to derive full conditional posterior distributions conditioned on $w$ than on $y$.

The conditional posterior distributions are provided by the following facts. Throughout, let $\theta | \cdot$ denote the full conditional posterior of parameter $\theta$.

**Fact 1:** The conditional posterior distribution of $w_{ij}$ is

$$w_{ij} | \cdot \overset{\text{indep.}}{\sim} \begin{cases} \text{TN}(c_{(y_{ij}-1)}, c_{(y_{ij})})(\mu^{(n)} + \alpha^{(n)}_i + \beta^{(n)}_j, 1), & s_{ij} = 0, \\ \text{TN}(c_{(y_{ij}-1)}, c_{(y_{ij})})(\mu^{(s)} + \alpha^{(s)}_i + \beta^{(s)}_j + \theta^{(s)} l_{ij}, \sigma^2), & s_{ij} = 1, \end{cases}$$

where $\text{TN}(a,b)(\mu, \sigma^2)$ is a Normal($\mu, \sigma^2$) distribution truncated below at $a$ and above at $b$.

**Fact 2:** Let $N_n = \sum_{i,j}(1-s_{ij})$ be the number of total trials in which nonstudied items are tested. Let $\lambda_n$ be the vector of grand mean, participant, and item effects for nonstudied trials: $\lambda_n = [\mu^{(n)}, (\alpha^{(n)})^T, (\beta^{(n)})^T]^T$. Let $w_n$ be the vector of $w_{ij}$ for nonstudied trials. Let $X_n$ be the $N_n \times (I + J + 1)$ design matrix such that $E[w_n] = X_n \lambda_n$. Let $1$ denote the covariance matrix of $w_n$. Finally, let $\Sigma^{(n)}_\lambda = \text{diag}(\sigma^2_\mu, \sigma^2_{\alpha^{(n)}}, \ldots, \sigma^2_{\beta^{(n)}}, \ldots)$. Then the full conditional posterior
distribution of \( \lambda_n \) is

\[
\lambda_n \mid \Sigma_w^{(n)}, \Sigma_\lambda^{(n)}, w_n \sim \text{MVNormal}_q \left( \frac{1}{\sigma_n^2} V_n X_n^T w_n, V_n \right),
\]

where \( q = I + J + 1 \) and \( V_n = \left( \frac{1}{\sigma_n^2} X_n^T X_n + (\Sigma_\lambda^{(n)})^{-1} \right)^{-1}. \)

**Fact 3:** Let \( N_s, \lambda_s, w_s, X_s, \) and \( \Sigma_\lambda^{(s)} \) be defined analogously to the comparable quantities in Fact 2. Then the full conditional posterior distribution of \( \lambda_s \) is

\[
\lambda_s \mid \Sigma_w^{(s)}, \Sigma_\lambda^{(s)}, w_s \sim \text{MVNormal}_q \left( \frac{1}{\sigma_s^2} V_s X_s^T w_s, V_s \right),
\]

where \( V_s = \left( \frac{1}{\sigma_s^2} X_s^T X_s + (\Sigma_\lambda^{(s)})^{-1} \right)^{-1}. \)

**Fact 4:** The full conditional posterior distributions of \( \sigma^2, \sigma_{\alpha(n)}^2, \sigma_{\beta(n)}^2, \sigma_{\alpha(s)}^2, \sigma_{\beta(s)}^2 \) are

\[
\sigma^2 \mid w_{ij}, \lambda \sim \text{IG} \left( a + \frac{1}{2} N_s, b + \frac{1}{2} \sum_{i,j} s_{ij} (w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \beta_j^{(s)} - \theta^{(s)} l_{ij})^2 \right),
\]

\[
\sigma_{\alpha(n)}^2 \mid w_{ij}, \lambda \sim \text{IG} \left( e + \frac{I}{2}, f + \frac{1}{2} \sum_i (\alpha_i^{(n)})^2 \right),
\]

\[
\sigma_{\alpha(s)}^2 \mid w_{ij}, \lambda \sim \text{IG} \left( e + \frac{I}{2}, f + \frac{1}{2} \sum_i (\alpha_i^{(s)})^2 \right),
\]

\[
\sigma_{\beta(n)}^2 \mid w_{ij}, \lambda \sim \text{IG} \left( e + \frac{J}{2}, f + \frac{1}{2} \sum_j (\beta_j^{(n)})^2 \right),
\]

\[
\sigma_{\beta(s)}^2 \mid w_{ij}, \lambda \sim \text{IG} \left( e + \frac{J}{2}, f + \frac{1}{2} \sum_j (\beta_j^{(s)})^2 \right).
\]

**Fact 5:** The full conditional posterior distribution of \( c_{i\ell} \) is
\(c_{i\ell} \mid \cdot \text{ indep.} \sim \text{Unif}\left(\max_j[w_{ij} \text{ such that } y_{ij} = \ell], \min_j[w_{ij} \text{ such that } y_{ij} = \ell + 1]\right), \ell = 2, \ldots, K - 2. \quad (2.11)\)

In Facts 2 and 3, separate full conditional posterior distributions were presented for \(\lambda_s\) and \(\lambda_n\) rather than for \(\lambda = (\lambda_s^T, \lambda_n^T)^T\). The reason for this separation is computational speed. The full conditional posteriors require inverting precision matrices and the speed of this operation is a power function of matrix rank. The precision matrices for the full conditional posteriors of \(\lambda_s\) and \(\lambda_n\) have half of the rows of the precision matrix of the full conditional posterior of \(\lambda\). Inverting the two, smaller matrices is significantly faster than inverting the larger one.

### 2.2.4 Sampling Algorithms

In Gibbs sampling, it is necessary to sample from the conditional posterior distributions. Fortunately, all of the conditional posterior distributions in Facts 1-5 are easy to sample from and details are provided in Rouder et al. (2007). Although sampling is straightforward, the resulting MCMC chains show a large degree of autocorrelation, especially in the criteria (see Figure 2.5, panels A and B). A high degree of autocorrelation is undesirable because obtaining adequate convergence requires a large number of Gibbs sampling iterations.

The source of the autocorrelation may be diagnosed. The distribution of \(c_{i\ell}\) in Eq. (2.11) is uniform between two values: the maximum \(w_{ij}\) in the response category below the criterion, and the minimum \(w_{ij}\) in the response category above the criterion. As shown in Figure 2.4, from iteration to iteration, the maximum latent value in response category \(k\) and minimum latent value in category \(k + 1\) will be close to one.
another, restricting the range of the full conditional posterior distribution. Because
the range of the distribution is restricted, samples of $c_{i\ell}$ will change very little from
iteration to iteration in the Gibbs sampler. In Figure 2.4, each response category
contains 20 responses; the problem gets worse with more responses in each category.
The result is high autocorrelation in the criteria chains, as shown in Figure 2.5,
panels A and B.

Figure 2.4: Source of autocorrelation in criteria chains. Panels from top to bottom
represent successive MCMC iterations. Dashed and dotted lines represent latent
variables $w_{ij}$ in categories $k$ and $k + 1$, respectively. The shaded box is a uniform
distribution from which criterion $k + 1$ is sampled.

In order to mitigate this autocorrelation, we can modify the sampling scheme.
Instead of sampling from the full conditional posterior of $c_{i\ell}$ in Fact 5, we sample
Figure 2.5: Top panels show segments of the MCMC chain for a selected criterion with Gibbs sampling. Bottom panels show autocorrelation functions for the chains.

from the conditional distribution

\[ \int [c_i, w | \lambda_n, \lambda_s, \sigma_n^2, \sigma_s^2, y] \, dw = [c_i | \lambda_n, \lambda_s, \sigma_n^2, \sigma_s^2, y] \]  

on every iteration of the Gibbs sampler. Integrating over the latent variables leads to more efficient MCMC chains (Holmes & Held, 2006). Let \( c^{(t)}_{i\ell} \) be a sample of the \( \ell \)th criterion for the \( i \)th participant, on the \( t \)th iteration of the MCMC chain. On each
iteration $t$, the following Metropolis-Hastings step is implemented to sample from (2.12).

**Step 1.** For each participant $i$, sample $K - 2$ independent values $z_{i2}^{(t)}, \ldots, z_{i(K-2)}^{(t)}$ from a Normal($0, \sigma_d^2$) distribution. The value $\sigma_d^2$ is chosen before analysis.

**Step 2.** Let $c_{i\ell}^{*(t)} = c_{i\ell}^{(t-1)} + z_{i\ell}^{(t)}$ for $\ell = 2, \ldots, K - 2$. The parameter $c_{i\ell}^{*(t)}$ serves as a proposal for a new sample of $c_{i\ell}$.

**Step 3.** For each participant, check that all proposal criteria are in the correct order, i.e. $c_{i\ell}^{*(t)} < c_{i(\ell+1)}^{*(t)}$ for $\ell = 1, \ldots, K - 1$. If the criteria are out of order for participant $i$, set $b_i = 0$ and skip Step 4 for participant $i$.

**Step 4.** For each participant, compute the likelihood of the model given the proposal criteria $c_{i\ell}^{*(t)}$, and the likelihood of the model given the samples $c_{i\ell}^{(t-1)}$, then compute its ratio, $b_i$:

$$
 b_i = \prod_{j=1}^{J} \left[ \Phi \left( \frac{c_{i(y_{ij})}^{*(t)} - d_{ij}^{(n)}}{\sqrt{\sigma^2}} \right) - \Phi \left( \frac{c_{i(y_{ij})}^{*(t)} - d_{ij}^{(n)}}{\sqrt{\sigma^2}} \right) \right]^{s_{ij}}
$$

$$
 \times \prod_{j=1}^{J} \left[ \Phi \left( \frac{c_{i(y_{ij})}^{(t-1)} - d_{ij}^{(n)}}{\sqrt{\sigma^2}} \right) - \Phi \left( \frac{c_{i(y_{ij})}^{(t-1)} - d_{ij}^{(n)}}{\sqrt{\sigma^2}} \right) \right]^{1-s_{ij}}
$$

(2.13)

**Step 5.** For each participant, accept the proposal criteria $c_{i\ell}^{*(t)}$ as the new sample for the criteria $c_{i\ell}^{(t)}$ with probability $\min(b_i, 1)$. Otherwise, let $c_{i\ell}^{(t)} = c_{i\ell}^{(t-1)}$.

Because the expression $b_i$ is not dependent on latent data $w$, this source of autocorrelation is eliminated. Figure 2.5 shows a sample of a chain for a selected criterion and its autocorrelation function. As can be seen, the new sampling scheme is highly effective.
The value of $\sigma_d^2$ must be chosen before the analysis. Some values of $\sigma_d^2$ are more effective than others. If the value is too high, values of $c_{it}^{*}$ are unreasonably large or small, and the candidate is rejected too often. If the value is too low, candidates do not deviate enough from the last sample and the chains will be highly autocorrelated. One option is to experiment with several values of $\sigma_d^2$ until an acceptance rate (Step 5) between 35% to 45% is achieved. Another option (which is invoked in the \texttt{hbmem} package) is to auto-tune $\sigma_d^2$ during the burnin period. That is, for the first several iterations increment $\sigma_d^2$ if the cumulative acceptance rate is too high, and decrement it if the acceptance rate is too low. If these first iterations are discarded, and $\sigma_d^2$ is held constant for the remaining iterations, then the chain will be an accurate sample from the posterior.

### 2.2.5 Model Performance

The hierarchical UVSD model is used here as a psychometric measurement tool. I performed several simulations to show that for this purpose, the model is robust. That is, it may be used to measure ROC asymmetry even if UVSD or the constituent hierarchical assumptions fail:

1. **How well does the model measure $\sigma$?** Data were simulated from the equal-variance signal detection model (with true values from estimates in the subsequently reported experiment) and fit with the unequal-variance model. Over 100 such simulations, the mean estimate of $\sigma$ was 1.04, and 95% of the estimates fell between .998 and 1.08. Although these simulations show that the estimate of $\sigma$ has a slight positive bias, it can distinguish between symmetric ROCs ($\sigma = 1$) and those typically observed ($\sigma = 1.28$). This slight bias is non-asymptotic, i.e., it diminishes with more participants or more items. The bias may be mitigated completely by reparameterizing the model such that both
the lowest and highest criteria are fixed (e.g., at 0 and 1, respectively) and the variances of both the new and studied-item distributions are free parameters. The ratio of the studied- to new-item strength variances are \( \sigma^2 \) in the current parameterization, and this ratio provides an unbiased estimate of \( \sigma^2 \). This model is developed fully in Morey, Pratte, and Rouder (2008); The estimate of \( \sigma^2 \) in Experiment 1 is the same for both parameterizations, and thus, does not reflect bias.

2. **What happens if effects are not distributed as normals?** The assumption of normally-distributed participant and item effects is not important for measuring \( \sigma^2 \). To show this fact, data were simulated in which participant and item effects were distributed as exponentials. Estimates of \( \sigma^2 \) for these simulations were as good as those from simulations with normally-distributed effects.

3. **What happens if the additivity assumption fails?** The effects of items, people, and lags are assumed to be additive; i.e., no interactions amongst these factors exist. This explicit treatment of participant and item variation is far more realistic than the implicit assumption made when data are averaged — that there are no item, participant, or lag effects. Nonetheless, potential violations of additivity may affect the estimates of \( \sigma \). The effects of violations in additivity are examined fully after discussing Experiment 1, as the results guide choices in simulation.

4. **What happens if the underlying UVSD model is wrong?** To show that the model accurately measures asymmetry in general, data were generated from Yonelinas’ (1994) dual-process model with participant and item effects on familiarity, and one recollection parameter \( (R) \) across all participants and items. I ran 50 simulations in which the true value of \( R \) ranged from 0 to .6 in increments of .012. Estimates of \( \sigma \) were obtained by fitting the UVSD model to these data.
The corresponding estimates of $\sigma$ are given approximately by $\hat{\sigma} = 1.02 + 2.8R$ (this linear relationship accounted for 97% of the variance). Thus, $\sigma$ serves as a reasonable index of ROC asymmetry even when this asymmetry arises from recollection rather than strength variance.

### 2.3 Model Restrictions and Comparisons

The goal on this chapter is to localize the source of ROC asymmetry. To do so, I construct a set of sub-models that represent theoretically important restrictions on the full hierarchical model developed above. The relationships among these models are shown in Figure 2.6. The top-left model, denoted $M_1$, is the full model presented above. The label “PIL$\sigma$” indicates that participants (P), items (I), and lags (L) have variable effects and that $\sigma$ is not constrained. The top-right model, denoted $M_2$ is the same as $M_1$ but with the restriction that $\sigma = 1$. Models $M_3$ and $M_4$ posit no effect of lag. Models $M_5$ and $M_6$ do not allow item or participant variability, respectively. Models $M_7$ and $M_8$ do not allow for any surface variation. Models $M_3$ through $M_6$ are referred to as *partial averaging* models as they are equivalent to averaging data over at least one surface variable. Models $M_7$ and $M_8$ are referred to as *full averaging* models as they are equivalent to averaging over people, items, and lags.

Given the set of models in Figure 2.6, the question is which one provides the best account of data. Assessing model fit is an active area of research, but the core of all methods involves assessing 1) how well the model accounts for the data, and 2) how flexible, or complex, the model is. For many methods such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), model fit is assessed as the *deviance*, which is defined as -2 times the log likelihood of the data given the model and it’s estimated parameters. Assessing model complexity is far less straightforward and characterizes the main difference between different model-fit
statistics. For example, the AIC statistic incorporates a penalty term that is two times the number of model parameters — the best-fitting model by AIC is the model that best explains the data with the fewest number of parameters. The BIC statistic incorporates a larger penalty for complexity, and this penalty is also a function of the number of parameters.

The models considered here are hierarchical, that is, the parameters are not independent of one another, but are treated as coming from a common parent distribution. The result is that there is constraint in the model that is not reflected in the number of parameters (see Section 2.2). This fact is made most obvious by considering that whereas adding hierarchical structure clearly adds constraint, it also adds to the number of parameters (one parameter for each effect variance). Because complexity in hierarchical models cannot be captured by the number of parameters, both AIC and BIC are inappropriate for hierarchical model comparison. Instead, the models are compared with the deviance information criterion (DIC, Spiegelhalter, Best, Carlin, & Linde, 2002). Like AIC and BIC, DIC is defined as the deviance, plus a penalty term for model complexity. For DIC, however, the penalty term for complexity incorporates constraint from hierarchical structures by penalizing by the number of effective parameters. If the hierarchical structure imposes no constraint (e.g., if there is so much data that extreme estimates are not pulled in) then the number of effective parameters equals the number of parameters in the model. As the hierarchical structures impose more constraint by shrinking parameter estimates toward one another, the number of effective parameters decreases. If the pooling is so great that the hierarchical structure makes all effect parameters be the same (e.g., \( \sigma_{\alpha(n)}^2 = 0 \)) then none of those parameters count toward the complexity penalty. DIC, like AIC, is measured on a log-scale. Differences of 10 or more in DIC are interpreted as very strong evidence for the model with the smaller DIC.
2.4 Experiment 1

Experiment 1 is a large-scale confidence-rating recognition memory experiment in which 97 participants were tested on 480 items. Although a 480-item test list is large, the same length was used by Glanzer et al. (1999) who obtained a typical sensitivity ($\hat{d}' = .94$) and variance ($\hat{\sigma} = 1.28$). If previous estimates of $\sigma > 1.0$ are confounded with participant or item variability, then the estimate of $\sigma$ from the hierarchical model
(\(M_1\)) is expected to be close to 1.0, and model \(M_2\) should fit equally as well as model \(M_1\). Alternatively, if recognition memory ROC curves are truly asymmetric, then the hierarchical model estimate of \(\sigma\) should be greater than 1.0.

### 2.4.1 Method

#### Participants

Ninety-seven University of Missouri students participated in Experiment 1 in return for credit toward a course requirement.

#### Stimuli

The word pool for the experiment consisted of 480 words from the MRC psycholinguistic database (Coltheart, 1981). Words were between four and nine letters and had a Kucera-Francis frequency of occurrence between 1 and 200 (Kucera & Francis, 1967). Study lists were constructed by randomly sampling 240 of these words and there was a separate study list per participant. Test lists were composed of all 480 words. Presentation order was randomized at study and at test and across all participants. A second word pool was used for practice items. These practice items did not overlap with the main word pool.

#### Procedure

Participants began with a practice session in which they studied five items and made confidence ratings to a subsequent 10-item test list of words. Following practice, participants were presented the study list. Each word was displayed in the center of the screen for 1850 ms, followed by a 250 ms blank period before the next word was presented. Participants were told that their memory for the words would be tested. To insure that participants attended to each word, they were required to read each
out loud; compliance was monitored. Following study, participants completed the test phase. Each item was presented on the screen and participants rated their confidence using the ratings: “sure new”, “believe new”, “guess new”, “guess studied”, “believe studied”, and “sure studied.”

2.5 Model Fits

2.5.1 Measuring Variance Components

All of the models in Figure 2.6 were fit to Experiment 1. The differences in DIC for each model compared with that for model $M_1$ are shown in the figure. As can be seen, all submodels are inferior to the full model $M_1$. This result indicates that not only do participants, items, and lags contribute substantial variability, but that $\sigma$ is assuredly greater than 1.0.

Figure 2.7A shows the estimates of overall sensitivity $d'$ from each model. The models with fixed $\sigma$ ($M_2$, $M_4$, $M_8$) lead to lower estimates of sensitivity $d'$ than those with free $\sigma$ ($M_1$, $M_3$, $M_5$, $M_6$, $M_7$). This pattern is expected because $d'$ is defined with reference to the new-item distribution. More interesting is the effect of averaging. First, the more averaging, the lower the estimate of $d'$. This result is also expected; Rouder and Lu (2005) note that averaging artifactually reduces sensitivity estimates (see also Wickelgren, 1968). Second, and more importantly, averaging leads to an underestimation of sampling error. For $M_1$, the estimate of $d'$ has a credible interval with a width of .28 (see error bar ‘i’). For $M_7$, in contrast, the credible interval width is .05 (see error bar ‘ii’). The former interval is accurate while the latter is wrong. The reason for this underestimation in confidence intervals with averaging is that systematic variation across items and participants includes correlations in performance, which, in turn, reduces the effective sample size (see
Clark, 1973; Rouder & Lu, 2005). Here, the reduction in standard error is severe and leads researchers to have far too much confidence in a distorted estimate.

Figure 2.7B shows the estimates of $\sigma$. First, all models that have free parameter $\sigma$ indicate an estimate well above 1.0, validating the large DIC increase for equal variance models in Figure 2.6. Second, averaging has only minimal effects, with increased averaging associated with marginally smaller estimates.

In the example of averaging over hard and easy items in Chapter 1, I speculated that averaging artifactually increases estimates of $\sigma$. This speculation is clearly wrong. Why did the hierarchical model not provide a smaller estimate of $\sigma$ than the averaging methods? The answer lies in a comparison of participant and item variability in mnemonic strength across the new and studied-item conditions. Estimates of this variability for participants ($\sigma_{\alpha,s}, \sigma_{\alpha,n}$) and items ($\sigma_{\beta,s}, \sigma_{\beta,n}$) are shown in Figure 2.7C (first four bars). Here we see that although there is substantial systematic item and participant variability, it is similar in size across the new and studied-item conditions. In the example in Chapter 1, in contrast, there was sizable variability for studied items and none for new ones. Averaging data from the studied-item condition conflates estimates of process variability in the studied condition with studied-item variability. Averaging data from the new-item condition conflates estimates of process variability in the new condition with new-item variability. If both of these item-variability sources are the same size, the degree of inflation is about the same in both conditions. Because the estimate of $\sigma$ reflects the ratio of studied-item strength variance to new-item strength variance, averaging the data from both conditions inflates both variances about equally, and their ratio $\sigma$ is approximately preserved. Critically, this pattern cannot be predicted a priori and could not have been discovered without the hierarchical model analysis.

In the preceding section on psychometric properties, I addressed the possibility of violations of the additivity of surface variation. The observed pattern, in which
Figure 2.7: Model analyses. A: Estimates of $d'$ from the eight models. B: Estimates of $\sigma$ from the five unequal-variance models. Darker colors in A & B denote greater levels of averaging. C: Estimates of standard deviations of participant effects ($\sigma_{\alpha,n}, \sigma_{\alpha,s}$) and item effects ($\sigma_{\beta,n}, \sigma_{\beta,s}$) from model $M_1$ (left), and item effects from model $M_9$ which does not include word-frequency effects (right). Error bars are 95% credible intervals.
there are equal amounts of item and participant variability across conditions, provides
guidance for assessment. The worst-case scenario is that violations of additivity occur
only for studied items and not at all for new items, as this pattern of variability leads
to an artifactual overestimation of $\sigma$. To assess the implications, data were simulated
from the equal-variance model $M_2$ with the addition of an interaction term to only
the studied-item distribution mean. This interaction was of the same magnitude as
the item effects themselves. Over 100 such simulations, the mean estimate of $\sigma$ from
model $M_1$ was 1.21, and 95% of estimates fell within 1.16 and 1.26. This estimate is
certainly too high, reflecting the fact that interactions enter only for studied items.
Even so, it is far lower than and cannot account for our finding of $\sigma = 1.36$. Thus,
even in this worst-case and highly implausible scenario, violations of additivity cannot
account for the degree of asymmetry present in ROCs.

2.5.2 The Structure of Lag, Participant, and Item Effects

Although model $M_1$ was designed to measure ROC asymmetry, it is also useful for
assessing lag, participant, and item effects. I consider first the effect of lag on the mean
of the studied-item distribution. The DIC value of 249 for model $M_3$ in Figure 2.6
indicates that lag does indeed have an effect. The estimated slope of the lag effect,
the amount of change in $d'(s)$ per one-item of lag, is $\hat{\gamma} = -0.001$. Although this number
is small, the effect of lag across all 717 levels on $d'$ is .717, which is quite large given
that the average sensitivity is $d' = 1.38$. Even though there is a lag effect, this effect
does not have an appreciable influence on the estimate of $\sigma$ (compare models $M_3$ and
$M_1$ in Figure 2.7B).

Participants and items also had significant effects on the new- and studied-item
distribution means. In model $M_1$, $\hat{d}'$ values for participants ranged from .12 to 2.44,
and those for items ranged from .21 to 3.60. Figure 2.8A shows a scatter plot in which
the participant effects in the studied condition are plotted as a function of those in
Figure 2.8: Participant and item effects. **A:** Participant effects in the studied condition plotted as a function of those in the new condition. The positive correlation is a response bias. **B:** Item effects in the studied condition plotted as a function of those in the new condition. The negative correlation is a mirror effect. The solid line represents the word-frequency effects obtained from model $M_9$ in which word-frequency is linearly regressed onto item effects. **C:** Item effects in the studied condition plotted as a function of those in the new condition after word frequency is partialed out. The mirror effect seen in panel B is greatly attenuated.

the new condition. Figure 2.8B shows the same for items. For participants, there is a positive relationship ($r = .30, t(95) = 3.11$) — people with higher strength in the new condition have higher strength in the studied condition.$^1$ As discussed previously, this positive correlation indicates that, in addition to differences in sensitivity $d'$, people primarily vary in overall response bias (see Figure 4.4). In contrast, items exhibit a mirror effect: items that have higher strength in the new condition tend to have lower strength in the studied condition ($r = -.22, t(478) = 4.82$).$^1$

The ability to estimate item effects without distortion from averaging over participants also allows for an examination of the effects of item characteristics. Here, I highlight the effect of word frequency on mnemonic strength. Item effects for the new and studied-item conditions are plotted as a function of word frequency in Figures 2.9A and 2.9B, respectively. As can be seen, there is a substantial mirror effect—

---

$^1$It is not clear how to assess the statistical significance of this correlation. The associated $t$-value is highly significant. Yet, interpreting this value assumes that the estimates are independent. In this case, they are not as they are connected by a hierarchical structure. The current priors, which assume independence, likely result in a bias toward attenuating correlations (see Rouder et al., 2007) so that the interpretation of the $t$-value is, if anything, conservative.
Figure 2.9: Word frequency effects. **A** & **B**: Item effects in the new and studied conditions, respectively, plotted as a function of word frequency on a log scale. Lines are non-parametric smooths (Cleveland, 1981).

Increased in word-frequency are associated with increases in baseline strength ($\beta_j^{(n)}$) and with decreases in studied strength ($\beta_j^{(s)}$).

The finding of strong word-frequency effects raises two additional questions: 1. Are there item effects in this data set above those from word-frequency? 2. If there are such extra-word-frequency item effects, do they also exhibit a mirror-effect pattern?

To answer these questions, a version of model $M_1$, termed model $M_9$, was constructed that includes word frequency as a covariate on item effects:

$$\beta_j^{(n)} \sim \text{Normal} \left( \phi^{(n)}(f_j - f_0), \omega_{\beta,n}^2 \right),$$

$$\beta_j^{(s)} \sim \text{Normal} \left( \phi^{(s)}(f_j - f_0), \omega_{\beta,s}^2 \right),$$

where $\phi^{(n)}$ and $\phi^{(s)}$ are slopes of the word frequency effect, $f_j$ is the logarithm of the Kucera-Francis word frequency for the $j$th item and $f_0$ is the mean of these log-word-frequencies such that the sum of these terms is zero. These parameters are estimated at $\phi^{(n)} = .15$ (CI$_{95} = [.14, .16]$), and $\phi^{(s)} = -.14$ (CI$_{95} = [-.15, -.12]$), which confirms the existence of a substantial frequency-based mirror effect.

The systematic item variability not accounted for by word frequency is given
by \( \omega_{\beta,n}^2 \) and \( \omega_{\beta,s}^2 \). These variances are useful for answering the first question about extra-word-frequency item effects. Estimates of these variances are shown as the two right-most bars in Figure 2.7C. The credible intervals are well above zero indicating substantial extra-word-frequency item effects. In fact, only 25% of the systematic item-level variance may be attributed to word-frequency; the remaining 75% is from other factors.

The second question about whether these extra-word-frequency item effects exhibit a mirror effect may be assessed by plotting item effects from model \( M_9 \) (see Figure 2.8C). Removing word-frequency effects greatly attenuates any item-level mirror effect (\( r = -0.007, t(478) = .15 \)). Whereas there is a mirror-effect for word-frequency, there does not appear to be one for other item characteristics that are orthogonal to word-frequency, even though these other characteristics account for a substantial proportion of item variability. We note, however, that this failure to find extra-word-frequency mirror effects may not generalize to other studies in which, for example, stimuli have a larger range on these orthogonal item characteristics (e.g., word length). The mirror effect has been interpreted as supporting a likelihood basis of decision making in which participants assess what their mnemonic strength would have been if the item was or was not studied (Glanzer et al., 1993). The current results indicate that participants may be able to make these calculations with regard to word frequency but not, perhaps, with regard to other item characteristics that are orthogonal to word-frequency.

### 2.6 Interpreting \( \sigma^2 \)

In the above analyses it was shown that, even in the absence of item variability, \( \hat{\sigma}^2 \) is greater than 1.0. In the context of the UVSD model, in which strength distributions are assumed to be normals, this result is interpreted to suggest that the variance of
mnemonic strength is greater for studied items than for new items. Unfortunately, however, the finding that \( \sigma^2 > 1.0 \) provides no evidence that the variance of latent mnemonic strength is different across the new and studied-item conditions unless one is willing to strictly (i.e., not approximately) assume that underlying strength distributions are normal. There is, however, no information to warrant this assumption. This fact has been known for some time (e.g., Egan, 1975), however, some researchers continue to champion the interpretation of \( \sigma^2 \) as strength variance (e.g., Wixted, 2007). In this section I show why this interpretation is not tenable.

Figure 2.10A shows an equal-variance normal signal detection model; Figure 2.10C shows the corresponding symmetric ROC curve. These normal distributions may be exponentially transformed, producing the distributions in Figure 2.10B. These distributions, termed log-normal distributions (Johnson, Kotz, & Balakrishnan, 1994), have the same shape but differ in their scales. Accordingly, the log-normal distributions in Figure 2.10B differ in both mean and variance. The ROC curve resulting from the log-normal distributions is also shown in Figure 2.10C. That is, the equal-variance normal model in Figure 2.10A and the unequal-variance log-normal model in Figure 2.10B produce exactly the same symmetric ROC curve.

The invariance of ROC curves to monotonic transformations of the underlying strength distributions holds generally. To illustrate why, consider the criterion in Figure 2.10A drawn at 1.0. The area above this criterion for new and studied items is 16% and 69%, respectively. These areas are the hit and false alarm rates for this criterion, shown as the point in Figure 2.10C. The criterion for the log-normal model in Figure 2.10B was generated by exponentiating the criterion from Figure 2.10A, that is, \( e^{1.0} = 2.72 \). The area above this criterion for the log-normal model for new and studied items is 16% and 69%, respectively. Because these areas are the same as for the normal model, the point on the ROC in Figure 2.10C also shows the hit and false alarm rates for the criterion in the log-normal model. As long as the same
Figure 2.10: ROC curves are invariant to transformations of the underlying strength distributions. **A.** An EVSD model. The vertical dashed line shows a criterion placed at 1.0 on the strength space. **B.** A signal detection model constructed by exponentiating the distributions in (A). These distributions are log-normal distributions with the same shape and varying scale. The vertical dashed line is the criterion from (A) on the exponentiated, exp(1). **C.** The ROC curve resulting from the signal detection models in both (A) and (B). The point corresponds to the criterion drawn in (A) and (B). **D.** A signal detection model in which strengths are distributed as gamma distributions with the same shape and varying scales. **E.** A signal detection constructed by taking the log of the distributions in (A). These distributions are log-gamma distributions with equal variances and varying means (shifts). **F.** The ROC curve resulting from the signal detection models in both (D) and (E).

transformation is applied to both the new and studied-item distributions, and the criterion placed on that strength space, the area above the transformed criterion under the transformed distributions is unchanged. Because there are an uncountable number of monotonic transformations, any given ROC curve is perfectly compatible with an uncountable number of signal detection models. The result is that ROC data provide no information regarding the shape of underlying strength distributions (e.g.,
normal vs. log-normal).

Figure 2.10D shows a signal detection model in which strengths are distributed as gamma distributions with the same shape and varying scales. Figure 2.10E shows a signal detection model that results from taking the logarithm of the gamma distributions in Figure 2.10D. These distributions differ only in shift, and therefore have the same variance. The resulting ROC curve for the signal detection models in Figure 2.10D and Figure 2.10E is shown in Figure 2.10F. This ROC curve is asymmetric, even though the distributions in Figure 2.10E have the same variance.

Figure 2.10 provides an example of an unequal-variance signal detection model that produces symmetric ROC curves (B), and an equal-variance model that produces asymmetric ROC curves (E). These signal detection models have non-normal strength distributions, and the figure clearly demonstrates that ROC asymmetry has nothing to do with the variance of underlying strength distributions if they are not strictly normal. Because ROC curves are incapable of providing evidence that strength distributions are normal, the finding that $\hat{\sigma}^2 > 1.0$ implies only that ROC asymmetry exists. The fact that asymmetry remains in the above hierarchical analysis implies that it is a feature of the mnemonic process, and not due to item variability.

There are several plausible models of recognition memory other than the UVSD model developed above that can account for ROC asymmetry. Such models include signal detection models that do not assume normal strength distributions, such as the gamma model in Figure 2.10D, and the Yonelinas dual-process model. Because these are non-linear models, in Chapter 3 I develop hierarchical versions of them that account for participant and item variability. The models are then compared via DIC.
Chapter 3

Hierarchical Dual- and Single-Process Models of Recognition Memory

In Chapter 2 a hierarchical version of UVSD was developed. Fitting this model to a typical experiment revealed that ROC asymmetry is not an artifact of averaging data over items, but rather, reflects a characteristic of the mnemonic system. The goal of Chapter 3 is to develop and compare hierarchical models that account for this asymmetry. Although a great deal of effort has been made to determine whether the single-process (UVSD) or dual-process (DPSD) views provide a better account of ROC data, no consensus has been reached. For example, in a recent review of the literature, Yonelinas and Parks (2007) conclude:

The authors discuss the methodological issues involved in conducting and analyzing ROC results. . . . The results indicate that there are at least 2 functionally distinct component/processes underlying recognition memory.

In another review published at the same time, and based largely on the same literature as that used by Yonelinas & Parks, Wixted (2007) concludes:
The unequal-variance signal-detection model and a dual-process threshold/detection model, accurately describe the receiver operating characteristic, . . . , and that literature is almost unanimous in its endorsement of signal-detection theory.

How could such disparate conclusions be reached? The most obvious reason is that the UVSD and DPSD models make very similar predictions for a given ROC curve. In order to observe curves that are fit better by one model or another, it is necessary to measure ROC curves across a range of accuracies. By using the hierarchical models, such a range is obtained naturally by taking participant and item variability into account. These models must account for how participants and items differ from one another, making them easier to distinguish. Moreover, in previous studies the model fits and fit statistics were distorted by averaging. This distortion is also avoided by comparing hierarchical models. For these reasons, in this chapter hierarchical versions of several single- and dual-process models are compared.

Because the purpose of the UVSD model developed in Chapter 2 was to provide a simple estimate of ROC asymmetry, a single value of $\sigma^2$ was assumed for all participants and items. In this chapter a UVSD model is developed in which each person and item has a unique value of $\sigma^2$. Fitting this model to the data from Experiment 1 reveals that $\sigma^2$ varies across people and items in a systematic way — sensitivity $d'$ and $\sigma^2$ are positively correlated across people and items. A gamma signal detection model is then developed that accounts for this relationship between sensitivity and asymmetry without the need to specify two separate parameters for each person and item. Following development of this true single-process model, the Yonelinas dual-process model is expanded to simultaneously account for person and item variability. These models are then compared using DIC.
3.1 Extended Hierarchical UVSD Model

In a typical UVSD analysis, the model is fit separately to each condition. In doing so, the majority of previous studies have found that conditions exhibiting higher sensitivity, \( d' \), also have higher values of \( \sigma^2 \). For example, in both meta-analysis and their own series of experiments, Glanzer et al. (1999) found that manipulations of study time, word frequency, encoding task at study, and several other variables affected both \( d' \) and \( \sigma^2 \). Likewise, in a meta-analysis Yonelinas and Parks (2007) found a positive correlation between \( d' \) and \( \sigma^2 \) across several different studies. Given the plausibility of this relationship between conditions and even experiments, it is reasonable to ask two questions: 1) Whether \( \sigma^2 \) varies across people and items, and 2) If so, whether there is a positive relationship between \( d' \) and \( \sigma^2 \) across people and items. To answer these questions, in this section an extended UVSD model is constructed that allows \( \sigma^2 \) to vary across participants and items. Experiment 1 is then analyzed with this extended UVSD model.

3.1.1 Model Specification

In the extended UVSD model each participant and item is allowed a unique value of \( \sigma^2 \) such that latent strengths \( X_{ij} \) are distributed as follows:

\[
X_{ij} \sim \begin{cases} 
\text{Normal}(d_{ij}^{(n)}, 1) & \text{new}, \\
\text{Normal}(d_{ij}^{(s)}, \sigma_{ij}^2) & \text{studied},
\end{cases}
\]

This extended UVSD model (E-UVSD) is identical to the UVSD model presented in Chapter 2 in that each person and item has a unique \( d_{ij}^{(n)} \) and \( d_{ij}^{(s)} \) that are additive combinations of a grand mean, a participant effect, an item effect, and a lag effect for \( d_{ij}^{(s)} \). Also, the middle criterion is fixed to zero and the remaining criteria are free to vary across participants. The model differs from that in Chapter 2 in that
the studied-item variance $\sigma^2$ is not fixed, but rather, it’s logarithm is the additive combination of a participant, item, and lag effect:

$$\log(\sigma^2_{ij}) = \mu^{(\sigma)} + \alpha_i^{(\sigma)} + \beta_j^{(\sigma)} + \theta^{(\sigma)} l_{ij},$$

where $\mu^{(\sigma)}$ is a grand mean, $\alpha_i^{(\sigma)}$ are participant effects, $\beta_j^{(\sigma)}$ are item effects, and $\theta^{(\sigma)}$ is the effect of the $l_{ij}$ lag. The additive model is placed on the logarithm of $\sigma^2$ to ensure that it is positive for all person by item combinations. As with the effects on $d_{ij}^{(n)}$ and $d_{ij}^{(s)}$, effects on $\sigma_{ij}^2$ are given hierarchical normal structures:

$$\alpha_i^{(\sigma)} \sim \text{Normal}(0, \sigma_{\alpha^{(\sigma)}}^2),$$
$$\beta_j^{(\sigma)} \sim \text{Normal}(0, \sigma_{\beta^{(\sigma)}}^2).$$

If all people have the same value of $\sigma^2$, then all estimates of $\alpha_i^{(\sigma)}$ will be near zero, as will $\sigma_{\alpha^{(\sigma)}}^2$. Likewise, if all items have the same value of $\sigma^2$ then $\sigma_{\beta^{(\sigma)}}^2$ will be near zero. Assessing the estimates of these effect variances will provide a test of whether it is necessary to allow people and items to have varying values of $\sigma^2$.

### 3.1.2 Model Analysis

Analysis of the extended UVSD model is similar to that for the UVSD model developed in Chapter 2. The only exceptions concern the posterior distributions of $d_{ij}^{(s)}$ and $\sigma_{ij}^2$. Studied-item latent data for the extended UVSD model is distributed as follows:

$$\omega_{ij}^{(s)} \sim \text{Normal}(d_{ij}^{(s)}, \sigma_{ij}^2),$$

and the goal is to sample the additive components in $d_{ij}^{(s)}$ and $\log(\sigma_{ij}^2)$ given these latent data. Let $\nu_{ij}^2 = \log(\sigma_{ij}^2)$. The full log conditional posterior distribution of the
additive components in $d_{ij}^{(s)}$ and $\nu_{ij}^2$ is

$$-\frac{1}{2} \left[ \sum_{i=0}^{I} \sum_{j=0}^{J} \left( \mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)} + \theta^{(s)} l_{ij} \right)^2 e^{(\mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)} + \theta^{(s)} l_{ij})} \right]$$

The grand mean and effects on $d_{ij}^{(s)}$ have the following normal conditional posterior distributions:

\[
\begin{align*}
\mathbb{E}[\mu^{(s)}|\ldots] &\sim N(m, v) \\
p &= \left( \sum_{i=0}^{I} \sum_{j=0}^{J} (e^{-\nu_{ij}^2}) + \frac{1}{\sigma_0^2} \right)^{-1} \\
m &= v \sum_{i=0}^{I} \sum_{j=0}^{J} \frac{w_{ij} - \alpha_i^{(s)} - \beta_j^{(s)} - \theta^{(s)} l_{ij}}{e^{\nu_{ij}^2}} \\
\mathbb{E}[\alpha_i^{(s)}|\ldots] &\sim N(m, v) \\
p &= \left( \sum_{i=0}^{I} \sum_{j=0}^{J} (e^{-\nu_{ij}^2}) + \frac{1}{\sigma_{a,s}^2} \right)^{-1} \\
m &= v \sum_{i=0}^{I} \sum_{j=0}^{J} \frac{w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \theta^{(s)} l_{ij}}{e^{\nu_{ij}^2}} \\
\mathbb{E}[\beta_j^{(s)}|\ldots] &\sim N(m, v) \\
p &= \left( \sum_{i=0}^{I} \sum_{j=0}^{J} (e^{-\nu_{ij}^2}) + \frac{1}{\sigma_{\beta,s}^2} \right)^{-1} \\
m &= v \sum_{i=0}^{I} \sum_{j=0}^{J} \frac{w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \theta^{(s)} l_{ij}}{e^{\nu_{ij}^2}} \\
\mathbb{E}[\theta^{(s)}|\ldots] &\sim N(m, v) \\
p &= \left( \sum_{i=0}^{I} \sum_{j=0}^{J} \left( l_{ij}^2 e^{-\nu_{ij}^2} \right) + \frac{1}{\sigma_0^2} \right)^{-1} \\
m &= v \sum_{i=0}^{I} \sum_{j=0}^{J} \frac{l_{ij} (w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \beta_j^{(s)})}{e^{\nu_{ij}^2}}
\end{align*}
\]
Conditional posteriors of the grand mean and effects on $\nu^2$ can be simplified from the full conditional, but do not have distributions with known forms and so are sampled independently with random-walk Metropolis-Hastings algorithms. To mitigate autocorrelation inherent in additive models, the effects on $d_{ij}^{(s)}$ and on $\nu_{ij}^2$ are given Metropolis-Hastings decorrelating steps.

### 3.1.3 Model Performance

To assess parameter recoverability for the extended UVSD model, data were generated from the E-UVSD model and fit with it. The design matrix, true grand means, criteria, and effect variances were taken from the E-UVSD analysis of Experiment 1, presented below. Figure 3.1 shows parameter estimates as a function of their true values. Clearly, the model is capable of recovering true parameters, including participant and item effects on $\sigma_{ij}^2$.

### 3.1.4 E-UVSD Fit to Experiment 1

The extended UVSD model was fit to the data from Experiment 1. There are two primary questions regarding this fit: 1) Does $\sigma^2$ vary across participants and items, and 2) If so, is there a relationship between these effects and those on sensitivity. These questions are answered in turn below.

**Variability in $\sigma^2$**

There are two (related) ways to assess whether participant and item effects on $\sigma^2$ are necessary. The first is by assessing model fit with DIC. The UVSD model in Chapter 2 with a single value of $\sigma^2$ across participants and items($M_1$) had a DIC value that was 872 points higher than the extended model presented above. That is, DIC overwhelming favors participant and item variability in $\sigma^2$. The extended
Figure 3.1: Parameter recovery of the E-UVSD model. Data were simulated from the extended UVSD model, and fit with it. All panels show estimated effects as a function of their true values. A. Participant effects on new-item mean. B. Participant effects on studied-item mean. C. Participant effects on log(σ²). D. Item effects on new-item mean. E. Item effects on studied-item mean. F. Item effects on log(σ²).

model has 1721 effective parameters, whereas the single-σ² model has only 1359 — DIC favors the more complex model as the gain in fit outweighs the gain in flexibility as assessed by the number of effective parameters.

As an alternative to DIC, we can examine the posterior distributions of effect variance; if effects are absent then these variances should be near 0. Figure 3.2 shows the posterior distributions of σ²_{α(s)} and σ²_{β(s)}, the size of participant and item effects on σ², respectively. That the mass of both distributions is substantially greater than 0.0 suggests that σ²_{α(s)} > 0 and σ²_{β(s)} > 0. Although both participant and item effects are needed on σ², the greater posterior distribution of σ²_{β(s)} over σ²_{α(s)} suggests that
there is more item variability in $\sigma^2$ than participant variability.

**Relationship between $\sigma^2$ and strength**

The top panels of Figure 3.3 show participant effects on $\sigma^2$ as a function of participant effects on sensitivity $d'$ (panel A), new-item strength $\alpha^{(n)}$ (panel B), and studied-item strength $\alpha^{(s)}$ (panel C). The bottom panels of Figure 3.3 show the same for items. Clearly there is a positive relationship between sensitivity $d'$ and $\sigma^2$ for both people and items. Examining this relationship between $\sigma^2$ and the new- and studied-item components reveals a negative relationship between $\sigma^2$ and new-item strength for both participants and items. These correlations are significant by Pearson correlation tests for both participants ($r=-.48$, $t=5.3$) and items ($r=-.64$, $t=18.1$). The relationship between $\sigma^2$ and studied-item strength is absent for participants ($r=-.12$, $t=1.2$), and positive for items ($r=.44$, $t=10.8$). Because sensitivity $d'$ is defined as $d^{(s)} - d^{(n)}$, these findings are perfectly in line with previous demonstrations that there is a positive correlation between $\sigma^2$ and $d'$ across conditions. The re-parametrized model used
here, however, reveals that this correlation is primarily driven by the relationship between $\sigma^2$ and new-item strength.

**Figure 3.3:** Relationship between $\sigma^2$ and strength. A. Participant effects on $\sigma^2$ as a function of sensitivity $d'$ defined as $d_i^{(s)} - d_i^{(n)}$. B. Participant effects on $\sigma^2$ as a function of those on new-item strength. C. Participant effects on $\sigma^2$ as a function of those on studied-item strength. D. Item effects on $\sigma^2$ as a function of sensitivity $d'$ defined as $d_j^{(s)} - d_j^{(n)}$. E. Item effects on $\sigma^2$ as a function of those on new-item strength. F. Item effects on $\sigma^2$ as a function of those on studied-item strength.

It is reasonable to wonder whether the patterns in Figure 3.3 reflect real data patterns, or artifacts of estimation. In the E-UVSD simulation presented above, participant and item effects were generated from independent normals, and so the true correlations between these effects on strength and $\sigma^2$ were zero. The estimated correlation between studied-item strength and $\sigma^2$ was .16 for participant effects, and .01 for item effects. Critically, the estimated correlation between new-item strength and $\sigma^2$ was -.06 for participant effects, and .03 for item effects. None of these correlations was significant, suggesting that the correlation found for the experimental data is not
an artifact of estimation.

There are several possible interpretations of this correlation. At the level of ROC curves, however, the implication is simply that as sensitivity increases, so does ROC asymmetry. Rather than allow both $d'$ and $\sigma^2$ to be free parameters for every participant and item to account for asymmetry, then, it should be possible to construct a signal detection model such that the relationship between sensitivity and ROC asymmetry is built in. Such a model, using gamma strength distributions, is constructed in the next section.

### 3.2 Gamma Signal Detection Model

Results of the extended UVSD model fit to Experiment 1 imply that ROC asymmetry is a function of sensitivity. The downside of the extended UVSD model is that it has lost any respectable resemblance to a single-process model. As Figure 1.5 in Chapter 1 shows, allowing strength variances to differ results in a loss of the structure in ROC space that makes EVSD such a parsimonious model. If people, items, and conditions all affect two parameters ($d'$ and $\sigma^2$), this increase in flexibility must imply more complexity than a 1-process model. In this section I show that there are single-process signal detection models that can predict a relationship between ROC asymmetry and $d'$ without the need to posit two parameters such as in UVSD. This feature is gained by forgoing the assumption that strength distributions are normal. Since, as I showed in Chapter 2, the normality assumption is completely unwarranted, studying other options provides a promising path toward a single-process model that accounts for variability in ROC asymmetry.

There are several choices of strength distribution that will provide a relationship between $d'$ and ROC asymmetry (see, e.g., DeCarlo, 1998). In the model developed here, latent strengths are distributed as gamma distributions rather than as normal
distributions. The effect of study is to increase the scale of the distribution; participants set criteria on the latent space as before. Because this is a non-linear model, I develop a hierarchical version that accounts for participant and item variability.

### 3.2.1 Model Specification

Figure 3.4A shows the hierarchical gamma signal detection (GSD) model. Latent mnemonic strengths are:

\[
X_{ij} \sim \begin{cases} 
\text{Gamma}(2, \lambda_{ij}^{(n)}) & \text{new}, \\
\text{Gamma}(2, \lambda_{ij}^{(s)}) & \text{studied}.
\end{cases}
\]

The shapes of both gamma distributions are fixed to 2.0. This value was chosen not for any deep theoretical reasons but simply because gamma distributions with this shape provide about the right degree of asymmetry. To fix the scale of the space, the middle criterion is fixed to 1.0, and, as in UVSD, the remaining criteria are free to vary across participants. The scales of the new and studied-item distributions, \(\lambda_{ij}^{(n)}\) and \(\lambda_{ij}^{(s)}\), respectively, are free to vary across participants and items. Sensitivity in the gamma model may be defined as the ratio of studied to new-item scales \(d'_{ij} = \lambda_{ij}^{(s)}/\lambda_{ij}^{(n)}\). Additive models are placed the log of scale parameters so that these scales are restricted to be positive:

\[
\log(\lambda_{ij}^{(n)}) = \mu^{(n)} + \alpha^{(n)}_i + \beta^{(n)}_j,
\]

\[
\log(\lambda_{ij}^{(s)}) = \mu^{(s)} + \alpha^{(s)}_i + \beta^{(s)}_j + \theta^{(s)}_{ij}.
\]

These random effects are given the same hierarchical structures used previously in UVSD.

Figure 3.4B shows GSD ROC curves for several levels of sensitivity. As can be seen,
Figure 3.4: The gamma signal detection model. A. The structure of the hierarchical GSD model. B. Predicted GSD ROC curves for several levels of sensitivity $d'$.

The curves produce a similarly structured field as EVSD (see Figure 1.3), however, they are asymmetric. If these lines were fit with UVSD, there would be a positive correlation between $d'$ and $\sigma^2$. Thus, the relationship between $d'$ and $\sigma^2$ seen in the E-UVSD analysis is built into GSD without any gain in complexity over EVSD.

It may seem that the gamma distribution is a rather arbitrary choice for a signal detection model. Indeed it is. This insight, however, applies to the normal as well — there is no real reason to chose the normal distribution over alternatives (see Egan, 1975; Lockhart & Murdock, 1970; Rouder, Pratte, & Morey, in press). The gamma is superior a priori to the normal on pragmatic grounds. It explains the asymmetry in the data while predicting constrained and orderly ROC fields. Therefore, it is a reasoned alternative worthy of study. Moreover, the gamma signal detection model, unlike UVSD, is clearly a single-process model — the effect of study is in one parameter. It would, therefore, serve as a more reasonable competitor to dual-process models than UVSD does.
3.2.2 Model Analysis

Latent data for GSD are posited exactly as in UVSD except that they are distributed as gamma distributions with shape 2:

\[ w_{ij} \sim \begin{cases} \text{Gamma}(2, \lambda_{ij}^{(n)}), & \text{New}, \\ \text{Gamma}(2, \lambda_{ij}^{(s)}), & \text{Studied}. \end{cases} \]

and with conditional posteriors that are truncated gammas rather than truncated normals. I present here the conditional log posterior distribution of the components for the studied-item scales. Those for the new-item scales are equivalent, and criteria are sampled as they are in UVSD.

\[
\sum_{i=0}^{I} \sum_{j=0}^{J} \left[ -w_{ij}^{(s)} e^{(\mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)} + \theta^{(s)} l_{ij})} - 2 \left( \mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)} + \theta^{(s)} l_{ij} \right) ight] - \frac{1}{2} \left( \frac{(\mu^{(s)})^2}{\sigma_0^2} + \frac{(\alpha_i^{(s)})^2}{\sigma_{\alpha,s}^2} + \frac{(\beta_j^{(s)})^2}{\sigma_{\beta,s}^2} + \frac{(\theta^{(s)})^2}{\sigma_0^2} \right)
\]

The conditional posteriors of participant and item effects on new and studied-item scales have no known form, and so are sampled with Metropolis-Hastings random walk algorithms, and effects are de-correlated using the same methods as for the UVSD effects. This model can be fit with the R package \texttt{hbmem} with the function \texttt{sampleGamma()}. The GSD model performs very well in simulation; as good as equal-variance normal signal detection. The fit of the GSD model to Experiment 1 is explored in the Model Comparison section below, following development of a hierarchical dual-process model.
3.3 Hierarchical Dual-Process Model

3.3.1 Model Specification

The hierarchical version of the Yonelinas’ dual-process signal detection model is shown in Figure 3.5. The structure of the EVSD component governing familiarity is identical to that developed in Chapter 2. As before, the means of the new and studied-item distributions are the additive combinations of participant and item effects, and all but the middle criteria are free to vary across participants. In the hierarchical version, recollection, $R_{ij}$, varies across participants, items, and lags. Recollection can not be estimated for every participant-by-item combination without constraint. Additive structures are placed on the probit transformation of recollection (i.e., the quantile function of the standard normal):

$$\Phi^{-1}(R_{ij}) = \mu^{(r)} + \alpha_i^{(r)} + \beta_j^{(r)} \theta^{(r)} I_{ij},$$

where $\mu^{(r)}$ is a grand mean, $\alpha_i^{(r)}$ are participant effects, $\beta_j^{(r)}$ are item effects, and $\theta^{(r)}$ is the linear effect of lag. As with UVSD, these participant and item effects are given normal parent distributions:

$$\alpha_i^{(r)} \sim \text{Normal}(0, \sigma_{\alpha^{(r)}}^2),$$

$$\beta_j^{(r)} \sim \text{Normal}(0, \sigma_{\beta^{(r)}}^2).$$
3.3.2 Model Analysis

According to the hierarchical dual-process model, the probability that the $i$th person makes the $k$th response to the $j$th item is:

$$
\Pr(y_{ij} = k \mid \text{new}) = \Phi(d_{ij}^{(n)} - c_{ik}) - \Phi(d_{ij}^{(n)} - c_{i(k-1)}) ,
$$
$$
\Pr(y_{ij} = k \in \{1, \ldots, K-1\} \mid \text{studied}) = (1 - R_{ij}) \left[ \Phi(d_{ij}^{(s)} - c_{ik}) - \Phi(d_{ij}^{(s)} - c_{i(k-1)}) \right] ,
$$
$$
\Pr(y_{ij} = K \mid \text{studied}) = R_{ij} + (1 - R_{ij}) \Phi(c_{i(K-1)} - d_{ij}^{(s)}) .
$$

As with UVSD, sampling from conditional posteriors is made easier by conditioning on latent data rather than the multinominal data. For the new-item condition, latent
data are sampled exactly as they are in the hierarchical EVSD model.

\[ \omega_{ij}^{(n)} | \ldots \sim TN(c_{i(y_{ij} - 1)}, c_{i(y_{ij})})(d_{ij}^{(n)}, 1). \]

In addition to sampling latent data for the signal detection components, latent data are sampled for recollection in the manner common for estimating linear models on binomial probabilities (Albert & Chib, 1995). These latent data, denoted \( \omega_{ij}^{(r)} \), are positive on trials for which recollection occurred and negative otherwise. On trials for which the response was not equal to \( K \) (i.e., not “sure studied”) recollection assuredly did not occur and so the response must have arisen from the signal detection component. For these trials the mapping between multinomial and latent data is as follows:

\[
\begin{align*}
\omega_{ij}^{(s)} | \ldots & \sim TN(c_{i(y_{ij} - 1)}, c_{i(y_{ij})})(d_{ij}^{(s)}, 1) \quad y_{ij} = K \\
\omega_{ij}^{(r)} | \ldots & \sim TN(-\infty, 0) (\phi^{-1}(R_{ij}), 1)
\end{align*}
\]

Responses to studied items equal to \( K \) (“sure studied”) could have arisen from recollection or from familiarity that was above the \( K - 1 \) criterion. If recollection occurred (i.e., \( \omega_{ij}^{(r)} \geq 0 \)) then we have no information about familiarity. Alternatively, if recollection did not occur (\( \omega_{ij}^{(r)} < 0 \)) then familiarity must have been above the \( K - 1 \) criterion:

\[
\begin{align*}
\omega_{ij}^{(s)} | \ldots & \sim N(d_{ij}^{(s)}, 1), \quad \omega_{ij}^{(r)} \geq 0 \quad y_{ij} = K \\
\omega_{ij}^{(s)} | \ldots & \sim TN(c_{i}, \infty) (d_{ij}^{(s)}, 1), \quad \omega_{ij}^{(r)} < 0
\end{align*}
\]

In a similar manner, if familiarity was above the highest criterion then we have no information about whether recollection occurred or did not. Alternatively, if familiarity is below the \( K - 1 \) criterion and yet the response is \( K \), then recollection must
have occurred:

\[
\begin{align*}
\omega_{ij}^{(r)} | \ldots & \sim N(\phi^{-1}(R_{ij}), 1), \quad \omega_{ij}^{(s)} \geq C_i(K-1) \\
\omega_{ij}^{(r)} | \ldots & \sim TN_{(0,\infty)}(\phi^{-1}(R_{ij}), 1), \quad \omega_{ij}^{(s)} < C_i(K-1)
\end{align*}
\]

\[
y_{ij} = K
\]

The latent data \(\omega_{ij}^{(n)}, \omega_{ij}^{(s)}, \text{ and } \omega_{ij}^{(r)}\) are normally distributed with unit variance and means \(d_{ij}^{(n)}\), \(d_{ij}^{(s)}\), and \(\phi^{-1}(R_{ij})\), respectively. Sampling the additive components conditioned on these normally-distributed data is straightforward. Criteria are sampled with a Metropolis-Hastings algorithm as they are in the UVSD model.

Like EVSD, UVSD, and GSD, the hierarchical dual-process model may be estimated with the \texttt{R} package \texttt{hbmem}. The function \texttt{dpsdSim()} generates data from the hierarchical DPSD model, and the function \texttt{dpsdSample()} fits the hierarchical DPSD model to data.

### 3.3.3 Model Performance

To assess performance of the hierarchical DPSD model, data were generated from the model with the design matrix, grand means, criteria, and effect variances the same as those from the fit of the model to Experiment 1 (below). Figure 3.6 shows estimates of effect parameters as a function of their true values. Clearly the model is highly capable of recovering parameter values.

### 3.3.4 DPSD Fit to Experiment 1

There are two critical questions regarding the fit of the hierarchical DPSD model to Experiment 1: 1) Does recollection vary across people and items, and 2) Is there a relationship between recollection and familiarity, as there was between \(d'\) and \(\sigma^2\) in the UVSD fit. To answer the first question, posterior distributions of \(\sigma^2_{\alpha(r)}\) and \(\sigma^2_{\beta(r)}\) are shown in Figure 3.7. All of the mass of both distributions lies substantially above...
Figure 3.6: Parameter recovery for the hierarchical dual-process signal detection model. Data were generated from the hierarchical DPSD model and these data were fit with the hierarchical DPSD model. Panels A, B, & C show estimated participant effects as a function of their true values for the new-item mean, studied-item mean, and recollection, respectively. Panels D, E, & F show the same for items. 

zero suggesting that there is both participant and item variability in recollection. Both distributions are also very similar, suggesting that the magnitude of participant variability in recollection is similar to that of item variability.

The second question concerns the relationship between familiarity and recollection. Figure 3.8 shows how recollection and familiarity covary. Figure 3.8A shows the scatter plot for participant effects on recollection as a function of participant effects on sensitivity, where the later is given by $\alpha_i^{(s)} - \alpha_i^{(n)}$. Figure 3.8D shows the same for items. As can be seen, there is a positive relationship. Figures 3.8B and 3.8E show the scatter plots for recollection and new-item or baseline familiarity; Figures 3.8C
Figure 3.7: Posterior distributions of participant effects on recollection, $\sigma^2_{\alpha(r)}$ (solid line) and item effects on recollection, $\sigma^2_{\beta(r)}$ (dashed line). That the mass of both distributions lies substantially above zero indicates that recollection varies across both participants and items.

and 3.8F show the same for studied-item familiarity. Almost all of the relationship between sensitivity $d'$ and recollection is reflected in baseline familiarity $d^{(n)}$. Items that pre-experimentally have less familiarity tend to be recollected at higher rates.

One explanation of the relationship between recollection and baseline familiarity is the obvious: items that pre-experimentally have less familiarity tend to be recollected at higher rates. An alternative explanation is that the model is mis-specified and the correlation reflects an un-modeled high-threshold component for new items, perhaps from a recollect-to-reject model (Rotello, Macmillan, & Van Tassel, 2000). Another explanation is that recognition memory is mediated by a single-process model, and the over-specification of the dual-process model manifests as correlations among parameters. As an analogy, consider reaction time distributions, which, when they differ, almost always differ in both mean and variance (E. J. Wagenmakers & Brown, 2007). A single-process account of this effect is a scale effect such as the Gamma distributions that differ in scale in Figure 3.4A — when the scale of the gamma dis-
Figure 3.8: Correlations between DPSD effects in recollection and familiarity. **A.** Participant effects in recollection as a function of those in sensitivity $d'$. **B.** Participant effects in recollection as a function of those in new-item (baseline) familiarity. The solid line is the estimated linear relationship from the DPSD model in which recollection is forced to be a linear function of baseline familiarity. **C.** Participant effects in recollection as a function of those in studied-item familiarity. Panels **D-F** show the same relationships for item effects.

distribution changes, both its mean and variance change. Alternatively, reaction time data may be fit with normal distributions that differ in both mean and variance. This 2-parameter solution is over-specified if the data were generated from a model such as gamma distributions, but this over-specification will manifest as correlations between the mean and variance parameters of the normal distributions. In a similar vein, the correlations between recollection and familiarity may imply that we are throwing two parameters at data that were generated from one. This possibility is explored further in the next section and in chapter 4.
3.4 Model Comparison

In this section the ability of UVSD, DPSD, and GSD models to account for the data in Experiment 1 are compared. One of the main goals of the model comparisons herein is discriminating between UVSD and DPSD. I assessed how well this can be performed with DIC via simulation. Data were generated by both hierarchical UVSD and DPSD models with the true values based on model fits to Experiment 1. Each simulated data set was fit with both UVSD and DPSD, and the DIC difference was calculated. Histograms of these differences are shown in Figure 3.9. Clearly the DIC statistic accurately discriminates between data generated from the two models. Although this approach is not as extensive as the bootstrap assessment offered by E. J. Wagenmakers, Ratcliff, Gomez, and Iverson (2004), it provides confidence that these models are indeed discriminable with sample sizes such as those in Experiment 1.

Figure 3.10 shows the twelve models that were fit to the data from Experiment 1. Table 1 shows DIC statistics for each model. The DIC value for each model is the difference between that model’s DIC value and the DIC value for the DPSD model developed above. The main comparison of interest is between the signal-detection models (left branch) and the dual-process models (right branch). Within each class of models there are several sub-models that may be considered (lines in Figure 3.10 denote these nested relationships). These comparisons are considered in turn below.

**DPSD is selected over alternatives**

Every DPSD model was preferred to its corresponding UVSD counterpart. This trend holds regardless of restriction. DPSD also outperformed the EVSD and gamma single-process alternatives.
Figure 3.9: Discriminating between UVSD and DPSD as generating model. Data were simulated from the hierarchical UVSD model (100 simulations) or the hierarchical DPSD model (100 simulations) with grand means and effect variances set to those estimated in the experiment. For each simulation the difference in DIC (∆DIC) between the UVSD and DPSD model fits was calculated such that negative values indicate that UVSD provides a better fit. The histogram of ∆DIC from these 200 simulations shows that when the data were generated from UVSD (white) DIC prefers the UVSD model; when the data were generated from DPSD (grey) DIC prefers the DPSD model.

The dependence of recollection on familiarity

The best-fitting model was the general DPSD model with random effects. As shown previously, the participant and item effects on recollection are highly correlated with those on familiarity for this model (see Figure 3.8). The existence of this relationship may imply that recognition memory is mediated by a single process, and the dual-process model is over-fitting the data. If this is the case, then it should be possible to construct a dual-process model that assumes a relationship between recollection and sensitivity.
Based on the relationship between recollection and baseline familiarity in Figure 3.8, a DPSD model was constructed in which recollection was a linear function of
baseline familiarity:

\[ \Phi^{-1}(R_{ij}) = \mu^{(n)} + \phi_\alpha \alpha_i^{(n)} + \phi_\beta \beta_j^{(n)}, \]

where \( \phi_\alpha \) and \( \phi_\beta \) are the linear slopes relating the new-item strength effects to the recollection effects for participants and items, respectively. The lines in the middle panels of Figure 3.8 are from this restricted model with estimated slopes of \( \phi_\alpha = -1.03 \) and \( \phi_\beta = -1.49 \). Although this linear model provides a good account of the relationship, the DIC value (see Table 1, restriction \( R \propto d^{(n)} \)) suggests that the decrement in fit is not offset by the gain in parsimony when compared to the general DPSD model. The conclusion is that although baseline familiarity and recollection are substantially related, a deterministic linear relationship is not sufficient.

**Random vs. Fixed Effects**

In the all of the models developed in Chapters 2 & 3, participant and item effects were treated as random effects, and their variances were estimated parameters. This hierarchical modeling approach is preferred as it allows generalization of effects to other samples of people and items, and provides for more constrained models. Another option, however, is to treat participant and item effects as fixed effects. To do so the variances of these effects (e.g., \( \sigma^2_{\beta(s)} \)) are fixed at some value. In this fixed-effects case, the normal distributions on effects serve as prior distributions, and their variances are chosen to be large so that they have negligible effects on parameter estimates. Although random-effects models are often assumed, I digress here to explore fixed-effects versions of UVSD and DPSD, and compare these with their random-effects counterparts to assess the effect of hierarchical pooling.

As can be seen in Table 1, models with random effects are selected over models with fixed effects. This trend holds for both DPSD and UVSD, and both when item
effects are included and when they are excluded. The DIC statistics reveal that the advantage of the random-effect models is not in fit. Naturally, fixed-effect models fit better as they are less constrained. The DIC penalty terms reveal, however, that the gain in parsimony for treating people and item effects as samples from parent distributions more than offsets the loss in fit.

To better explicate the constraint in random effects modeling, participant and item effect estimates from the random-effects model are plotted as a function of those from the fixed-effects model. Figure 3.11 shows the case for the UVSD model, and the displayed trends are similar for DPSD. The main difference between random and fixed effects is at the extremes — random effect estimates are less extreme than their fixed effect counterparts. This trend reflects the natural correlate of assuming these effects are from a common parent distribution with normal tails. The interpretation from the random effects model is that the extremes in fixed-effect estimates are elements of over-fitting. The constraint in random effects is expressed as a shrinkage of extreme estimates to group means. This shrinkage effect is greatest for item estimates (bottom row) because there are many more items than participants. The shrinkage is also greater for studied-item parameters than new-item parameters because the effect of study is modeled in two parameters \(d(s), \sigma^2\) rather than in one \(d(n)\).

**The presence of item effects**

In previous sections I showed that according to DIC, there is substantial participant and item variability in both \(\sigma^2\) in UVSD and in recollection in the dual-process model. Two important questions remain concerning item effects: 1) should item effects be included in the models at all, and 2) is accounting for item effects driving the advantage of DPSD over UVSD? To answer both questions fixed- and random-effect UVSD and DPSD models without any item effects or lag effects were fit to Experiment 1.
The typical approach to fitting the UVSD and DPSD models is to average data over items to produce participant-specific effects. Averaging implicitly assumes that there are no item or lag effects. To mimic averaging, I implemented DPSD with item and lag effects in $d^{(n)}_{ij}$, $d^{(s)}_{ij}$, and $R_{ij}$ set to 0. The fixed-effects version of this no-item effects restriction is the closest Bayesian analog of the typical analysis, and is shown in Figure 3.10 as the box with hatched lines.

These analyses, which are comparable to averaging data over items, are shown in the last four rows of Table 1. Item effects are so prevalent that models without
item effects perform worse than EVSD (with item effects). This last comparison is important because EVSD has been considered insufficient for recognition memory for over three decades.

Results for the models without item effects make it evident that the inclusion of item effects is not driving the selection of DPSD over UVSD. Consider the comparison of the general models vs. the comparison of models without item effects. For the former, the DIC favors DPSD by 137; for the later, DIC favors DPSD by 44. The difference in these comparisons highlights that one of the advantages of modeling items is that it facilitates model selection. When accounting for items, the data are more rich (greater numbers of degrees of freedom) and the models are more complex. The balance, however, is that model mis-specification of structural properties becomes more apparent and so model selection is easier.

3.5 Conclusion

The results show some of the advantages of taking into account item and participant effects while modeling recognition memory performance. The substantive conclusion offered here is that the dual-process signal detection model outperforms the others. This result, however, should not prompt an overwhelming acceptance of DPSD. The following caveats are offered to aid in the judicious interpretation of the results.

The focus of this chapter has been on statistical modeling rather than on exploring the best manipulations to test theoretical positions. The DPSD model would be more convincing if one could selectively influence key parameters, especially while accounting for participant and item variation. Several researchers have made convincing arguments that previous demonstrations of selective influence are methodologically or conceptually flawed (Dunn, 2008; Wixted, 2007; Yonelinas & Parks, 2007). In Chapter 4, the hierarchical models developed above are used to offer a principled and
A second caveat is rooted in the substantial relationships in DPSD parameter estimates. Recollection is correlated with baseline familiarity, suggesting that familiarity and recollection may not be that distinct. There are several possible reasons for this dependence that retain the spirit of the dual-process model. For example, the correlation may reflect a mis specification of modeling recollection as a single high-threshold process; instead the pattern may indicate a recollect-to-reject component. Alternatively, the correlation may be diagnostic of a single-process structure. The challenge from this point of view is to explain why the model with the linear constraint fared worse than the general model. One possibility is that the wrong transforms were chosen for regression. In the DPSD model, recollection parameters enter through a probit transform and familiarity is assumed to be normally distributed. Both of these choices are arbitrary and perhaps other choices would reduce the noise in the scatter plots of Figure 3.11. The ability to detect mis-specification in transform increases as the data span larger ranges in accuracy. Though these ranges are large in our experiments, they may be made even larger through manipulation. In Chapter 4, such larger ranges of accuracy are obtained.

A third caveat reflects the nature of DIC as a model selection statistic. DIC has as a benefit that it reflects constraint from hierarchical structure. One downside is that it does not reflect this constraint in a consistent manner. As discussed in Chapter 2, DIC does not appropriately penalize complexity as sample size increases. DIC is most useful for selecting across models of relatively comparable complexity, such as between comparable UVSD and DPSD models. The DIC evaluation of nested models with large differences in effective parameters and with large sample sizes, as we have here, are cause for concern. As noted, DIC is based on the same logic as AIC, and as a consequence, overstates the evidence against nested submodels. In this regard, the selection of the general DPSD model over the model where recollection is a linear
function of familiarity should be taken with qualification. Similarly, the rejection of
the gamma model is qualified via DIC—it too may reflect DIC’s bias toward more
complex models in cases with large sample sizes. Ideally, more consistent model
selection techniques, such as Bayes factor or minimum description length (Grunwald,
Myung, & Pitt, 2005), would be available for this inference.
Chapter 4

Selective Influence of Dual-Process Parameters

In Chapter 3 the dual-process model was shown to provide a superior fit to recognition memory data over competing signal detection models by DIC. This result implies that DPSD outperforms selected competing models — it does not necessarily imply that DPSD is a good model of recognition memory. A more desirable test is to examine the performance of the DPSD model in its own right, rather than compare it with other models that are also likely wrong in some ways. To do so, in this chapter I explore whether the parameters of the DPSD model may be selectively influenced. In a Selective influence test the research manipulates some variable. The to-be-tested model makes predictions regarding which parameters should be affected by the manipulation, and which should remain constant across levels of the manipulation. If the parameters of the model are able to be selectively influenced, this result would provide strong evidence in favor of the model. Alternatively, if all manipulations only affect one parameter, or affect both parameters simultaneously, then the model is mis-specified.

For the dual-process model, the critical question is whether some manipulations
affect familiarity alone, whereas others affect only recollection. In Chapter 3 I showed that the effects of people and items exist in both recollection and familiarity. Moreover, these effects are correlated. There has been a good deal of work that has purportedly shown, however, that some manipulations are capable of influencing only recollection or only familiarity. Here I review this work, and discuss several critical flaws that render it questionable. Then, eight experiments are conducted and analyzed with the hierarchical DPSD model in an effort to obtain selective influence of dual-process model parameters.

4.1 Previous Claims of Selective Influence

Yonelinas (2002) conducted a meta-analysis to explore whether parameters of the dual-process model could be selectively influenced. His results are shown in Figure 4.1. The top panel shows the effects of encoding manipulations on familiarity and recollection estimates. The bottom panel shows the same for retrieval manipulations. Overall, these results imply that most manipulations affect both recollection and familiarity. When both parameters are affected, neither has been selectively influenced and such demonstrations do not provide strong evidence for the model. There are, however, some manipulations that may selectively influence recollection and familiarity. First, the effect of slow verses fast response deadline appears to affect only recollection. For slow deadline conditions participants take as much time as needed to make a recognition judgment; in speeded conditions they are forced to respond within a short (e.g., 1 second) time window after the test word is presented. This finding fits well within the dual-process theory — recollection takes time to develop whereas familiarity is automatic and develops quickly. When participants are forced to respond quickly, there is no time for recollection to develop. On the other hand, the effect of perceptual match vs. mismatch appears to be only in familiarity. An
example of a match-mismatch effect is when the fonts of items at test either match or mismatch the fonts that the words were originally studied in. According to the dual-process model, familiarity processes are based on perceptual characteristics being reinstated at study, whereas recollection processes involve higher-order characteristics such as semantic information. Thus, the match-mismatch effect in familiarity alone and the response-deadline effect in recollection alone imply that parameters of the dual-process model may be selectively influenced as predicted. There are, however, several problems with the studies used in this meta-analysis. These problems are considered in turn below.

4.1.1 Reliance on Averaged Data

All estimates in Figure 4.1 were obtained by averaging data over participants, items, or both. There is no way to know the extent of how this averaging may have distorted the results without knowledge of the true generating model and its parameters. One resulting distortion that is likely, however, it that confidence intervals on parameter estimates are biased to be too small, as shown in Chapter 2. When the test of a model is to show an effect of a manipulation in one parameter, and a lack of effect in another, it is critical that confidence intervals on parameters and the corresponding significant tests are accurate. In this case, with confidence intervals that are too small, it is likely that small effects of manipulations (see below for examples) had no reliable effect on either recollection or familiarity. However, because of distorted confidence intervals it is concluded that one (or both) parameters was affected. The only solution to obtaining accurate estimates and confidence intervals is to account for participant and item variability simultaneously as is done below.
Figure 4.1: Meta-analysis results from Yonelinas (2002). The top panel shows the average effects on recollection and familiarity parameters for several encoding manipulations. The bottom panel shows the same for retrieval manipulations. Parameter estimates are averaged over confidence ratings, remember-know, and process-dissociation procedure experiments.
4.1.2 Reliance on Remember-Know Paradigm

The majority of studies used in the Yonelinas (2002) review utilize the remember-know paradigm. In particular, the demonstration that perceptual match-mismatch affects only familiarity comes from two remember-know studies by Gregg and Gardiner (1994). In a remember-know task, instead of confidence ratings participants are given the response options “new”, “know”, and “remember”. They are instructed to respond “remember” if they remember studying the word, and to respond “know” if they know they studied the word but do not remember doing so. The proportion of “know” responses to studied items is interpreted as a measure of familiarity, the proportion of “remember” responses to studied items is interpreted as a measure of recollection. In Gregg and Gardiner, the effect of perceptual match-mismatch was to increase “know” responses, while leaving “remember” responses relatively unaffected.

Several critiques have been leveled against the remember-know paradigm (see Dunn, 2004, for a review). Many of these critiques concern the fact that effects on remember and know rates may reflect changes in response bias, rather than changes in two separate mnemonic processes (such as recollection and familiarity). For example, in the Gregg and Gardiner (1994) experiments, participants studied a list of words presented visually on the screen. At test, half of the words were presented visually (perceptual match) and half were presented auditorially (perceptual mismatch). The resulting remember and know rates are shown in Table 4.1. “Remember” responses to studied items were low, but relatively stable across the mismatch and match conditions; “know” responses to studied items doubled across mismatch and match conditions. Critically, however, “know” responses to new items also doubled across conditions. This effect is not surprising, as the match and mismatch conditions are confounded with visual and auditory presentation at test. It seems that when participants are presented with an auditory item they are more likely to respond that they studied it (“know” responses) regardless of whether it was actually studied or
Table 4.1: Results From Gregg & Gardiner

<table>
<thead>
<tr>
<th></th>
<th>Remember</th>
<th>Know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditory</td>
<td>.10</td>
<td>.27</td>
</tr>
<tr>
<td>Visual</td>
<td>.11</td>
<td>.52</td>
</tr>
<tr>
<td>New</td>
<td>.05</td>
<td>.09</td>
</tr>
<tr>
<td>Auditory</td>
<td>.09</td>
<td>.18</td>
</tr>
</tbody>
</table>

not. In this experiment, such a response-bias effect provides a perfectly viable explanation of the results without any recourse to recollection and familiarity. In fact, Dunn (2004) was able to fit these data and data from 400 other conditions extremely well with an equal-variance signal detection model designed for remember-know data.

4.1.3 Lack of Selective Influence in Confidence Ratings

Unlike remember-know, in the confidence-rating paradigm mnemonic processes such as recollection and familiarity, assuming they exist, are capable of being assessed independently of response biases. The question is, then, whether there exist convincing selective influence results in any ROC studies. One of the most impressive claims of selective influence in ROC curves comes from Yonelinas et al. (2002), who administered a confidence-rating task to normal participants, and to patients with medial temporal-lobe damage. Because the medial temporal lobe, including the hippocampus, is thought to underlie the recollection process, Yonelinas et al. hypothesized that only recollection estimates from DPSD should differ across controls and patients. The resulting ROC data are shown in Figure 4.2A. Fits of the dual-process model to these data imply a small effect of this brain damage on familiarity (.36 to .29), and a pronounced effect on recollection (.35 to .08).

The ROC curves for controls and brain-damaged patients differ in an important way that does not necessarily have anything to do with recollection or familiarity: Performance patients is worse than that for controls. As overall performance gets worse, the ROC curve approaches the diagonal. Accordingly, the small change in “fa-
Figure 4.2: A. ROC curves from controls (highest curve) and two groups of brain-damaged patients (lower two curves) in Yonelinas et al. (2002). Lines are predictions from the best-fitting gamma signal detection model presented in Chapter 3. B. ROC curves from all confidence-rating studies reviewed by Yonelinas (2002). C. Estimates of recollection from the ROC curves in B as a function of estimates of familiarity. Lines connect two conditions within a study. Colors represent different manipulations.

familiarity” accompanied by a large change in “recollection” may simply be an artifact of the ROC curve becoming more symmetric as overall performance worsens. The relationship between overall performance and asymmetry does not imply a two-process model. In fact, the lines in Figure 4.2A are from the single-process gamma signal detection model (see Chapter 3), in which ROC asymmetry and overall performance naturally co-vary. As can be seen, this single-process model provides a very good account of the ROC data. The result is that these data provide evidence of selective influence evidence in the dual-process model, only if perfectly viable single-process accounts of the data are ignored.

Figure 4.2B shows the ROC curves used in the meta-analysis performed by Yonelinas (2002). Strong evidence for the dual-process model would manifest as, for example, some straight lines (low familiarity) crossing over the curved ones. There is noise in these curves, however, overall they do not cross, but rather, exhibit the degree of curvature and asymmetry predicted by the gamma model: more asymmetry as overall performance increases. Figure 4.2C shows estimates of recollection as a function of estimates of familiarity from the ROC curves in Figure 4.2B. Although these data
were used as evidence for selective influence, it is clear that every manipulation considered in the review affected both recollection and familiarity to some extent. Thus, neither parameter was capable of being manipulated independently.

The plot of recollection as a function of familiarity in Figure 4.2C provides a convenient way of assessing selective influence, and more generally whether two processes are necessary to account for the data. Recall that if recollection and familiarity always co-vary, then this redundancy implies a single process. If this is the case, then the two-dimensional recollection-familiarity space should be able to be reduced to a single dimension by drawing a monotonically increasing line through all of the data points (taking noise into account). If such a line cannot account for the data, then this failure to reduce the space to one dimension implies that two processes are necessary. This analysis is similar to state-trace analysis (Bamber, 1979), but relies on plotting parameters rather than data.

The data in Figure 4.2C all lie roughly on a common trajectory that is similar to the diagonal. Without confidence intervals on the estimates, however, it is impossible to test whether violations of this uni-dimensional trajectory exist. Moreover, had confidence intervals been available they would have been obtained from averaged data, and thus, underestimated. A main goal of this chapter is to determine whether accurately estimated estimates of recollection and familiarity all lie on a single trajectory in this space, or if violations can be found implying a two-process model.

4.1.4 Small Effects

In order to provide a convincing demonstration of selective influence, it is necessary that the overall effect of a manipulation be substantial. If it is not, then a failure to find effects in one parameter may simply be due to a lack of power resulting from a small overall effect to begin with. One example is the match-mismatch effect in Figure 4.1. Although these results suggest that the effect of perceptual match-mismatch
is only in familiarity, the overall effect size is small compared to that for other mani-
pulations. It is therefore possible that, had a sizable effect been found, estimates of recollection would also be affected. In fact, the mere existence of the perceptual match-mismatch effect has been debated (Mulligan, Besken, & Peterson, 2010; Mulli-
gan & Hirshman, 1995), and no study has been conducted with large match-mismatch effects, much less large effects that affect only familiarity. In subsequent experiments I attempt to obtain a large match-mismatch effect in order to test whether such effects manifest in recollection as the overall effect increases.

The small match-mismatch effect is an example of a more general failure of re-
searchers to acknowledge that small effects are not useful for showing selective influ-
ence. For example, in a recent study Howard, Bessette-Symons, Zhang, and Hoyer (2006) claim to have provided evidence from ROC curves that aging selectively im-
pairs recollection, leaving familiarity relatively intact. ROC curves for their old and young participants are shown in Figure 4.3. From the ROC curves it is obvious that Howard et al. found little to no effect of aging on overall performance. I have taken the data from this study and converted the confidence ratings into simple binomial hit rates (any “studied” response to studied items) and false alarm rates (any “new” response to studied items). These rates were converted to $d'$ values for each participant by the standard EVSD definition ($\Phi^{-1}(hit) - \Phi^{-1}(fa)$). These binomial $d'$ values provide a very simple measure of overall recognition performance for each participant. A t-test on these $d'$ values reveals that Howard et al. found no effect of young vs. old adults on recognition memory performance by this crude binomial measure ($d_{young} = .9793, d_{old} = .9790, t(75) = .003, p \approx 1$). Nonetheless, in fitting the DPSD model and comparing estimates across the old and young groups, familiarity estimates were larger for the old than the young participants, but this difference was not significant ($t(69.4) = 1.41, p < .15$). The effect in recollection was in the op-
posite direction (young higher than old), and this effect was significant by a t-test
There are several reasons why the data from Howard et al. should not be taken as evidence of selective influence. One problem with this analysis is simply that the overall effect is so small that, if older adults were slightly impaired in familiarity there would be no chance of measuring such an effect in this study. Another problem concerns their conclusion that there is an effect on recollection. For one, because familiarity estimates were actually lower for young than old, the reverse pattern in recollection may simply reflect a trade-off between recollection and familiarity estimates when the model is applied to two conditions that do not differ at all. Such a trade-off between recollection and familiarity estimates is common, as the joint likelihood function contains a very shallow ridge. Moreover, because the confidence in these estimates are biased due to averaging over items, any marginally significant effect must be interpreted with caution. Even in the absence of biased confidence intervals, it is well-known that null hypothesis tests are biased to reject the null hypothesis (Edwards, Lindman, & Savage, 1963; Goodman, 1999; Jeffreys, 1961; Sellke, Bayarri, & Berger, 2001; E.-J. Wagenmakers & Grunwald, 2006). A more balanced assessment may be made using a Bayes Factor test (see Rouder, Speckman, Sun, Morey, & Iverson, 2009). This test reveals that for these data, the alternative hypothesis that recollection is different across old and young is only 1.6 times more likely than the null hypothesis that they are the same. Such a small Bayes Factor is considered marginal to zero evidence for a difference. The confluence of problems with the interpretation of Howard et al. are likely common. Unfortunately, such studies are often treated as strong evidence for the dual-process theory. The result is a collection of studies that looks like an impressive foundation for dual-process (e.g.,

\[ t(72.9) = 2.22, \ p < .03 \].

\(^1\)The Bayes Factor was derived from the reported t-value, which was obtained by assuming unequal variances in mean recollection across old and young. The t-test derived by Rouder et al. (2009) assumes equal variances. If Howard et al. had run a t-test assuming equal-variances, the t-value may have been somewhat larger. The difference is likely small, however, as the true degrees of freedom (77) and the corrected degrees of freedom reported (72.9) were similar.
Figure 4.3: ROC curves from young and old participants in Howard et al. (2006). Estimates of recollection and familiarity are those derived by Howard et al. (Figure 4.1). This foundation, however, falls apart when a closer look is taken at the piecewise evidence.

The goal of Experiments 3-6 is to determine whether parameters of the dual-process model can be selectively influenced. Manipulations of levels-of-processing, perceptual match-mismatch, and response deadline are considered. Before assessing selective influence of recollection and familiarity, however, I explore whether the criteria parameters in the dual-process model may be selectively influenced by a response bias manipulation. According to the model, changes in response biases for certain confidence-rating responses over others should effect the criteria, but not effect either recollection or familiarity estimates. If response bias manipulations can effect familiarity or recollection, then clearly these parameters are not measuring what they are intended to.
4.2 Experiment 2: Biasing Confidence Ratings

In the Yonelinas dual-process model, the probability of recollection is greatly determined by the probability of responding “sure studied”. It is therefore important that changes in the probability of making this response that are not due to recollection differences do not effect the estimate of recollection. In Experiment 2 the probabilities of making particular confidence rating responses (e.g., “sure studied”) were manipulated across conditions by using payoffs to bias participants toward certain responses over others. In the dual-process model the effect of this bias manipulation should manifest in criteria. The bias manipulation should not, however, have any effect on recollection or familiarity.

4.2.1 Methods

Participants

Twenty-two University of Missouri students participated in Experiment 2 in return for credit toward a course requirement.

Stimuli

Stimuli were the same 480 words used in Experiment 1.

Procedure

Participants began with a practice session in which they studied five items. Following study they preformed a practice test session. At test the screen displayed a running tally of how many points the participant had earned thus far (starting at 0), and a figure informing the participant of how many points each response would earn or lose, depending on whether the response was correct or incorrect, respectively. Figure 4.4A
Figure 4.4: Experiment 2 biasing conditions. Either panel A or panel B were displayed to participants throughout Experiment 2. Bars above zero indicate points gained if the new/studied response is correct, and those below zero indicate points lost if the new/studied response is incorrect. **A.** Point structure should bias participants to use the outside (“sure”) responses. **B.** Point structure should bias participants to use the inside (“believe”) responses.

shows the figure displayed at practice. Following a button press, a test word appeared and participants made one of the four available response options: “sure new”, “believe new”, “believe studied”, and “sure studied”. Following response, the point tally was updated, and participants were instructed to press a button to see the next word. Four response options were used instead of six because a pilot experiment suggested that participants were not able to successfully use the more complex 6-button point structures to bias responses.

Following practice participants completed the study phase by reading each of 240 randomly-selected items aloud. Following study, participants were instructed to carefully examine the point-structure graphic shown in Figure 4.4A in order to maximize
points. This point structure is termed *biased out* as it should bias participants to use the outer ("sure") responses. This point structure was kept on the screen for the first 120 test trials. After the 120th trial, a screen appeared to warn participants that the point structure was changing, and that they should examine the new graphic carefully. For the next 240 trials the point-structure graphic in Figure 4.4B was displayed on the screen. This point structure is termed *biased in* as it should bias participants to use the inside ("believe") buttons. Following these trials, the point structure changed back to that in Figure 4.4A for the remaining 120 trials. This ABBA design for bias conditions was employed for all participants.

### 4.2.2 Results

ROC curves for the biased-in and biased-out conditions, averaged over participants and items, are shown in Figure 4.5A. Even though these ROC curves are distorted due to averaging, they suggest that the bias manipulation was remarkably confined to moving hits and false alarms along a common trajectory. That is, the bias manipulation had no effect on the shape or position of the ROC curve, but had a substantial effect on where the data points lie upon it.

The hierarchical DPSD model was fit to data from the biased-out and biased-in conditions separately. In Figure 4.5B estimates of criteria from the biased-in condition are plotted as a function of those from the biased-out condition for each participant. In the biased-out condition criteria were close to zero, reflecting a high probability of using the outside "sure" responses. Criteria in the biased-in condition are further from zero, reflecting a higher probability of using the inside "believe" responses. Figure 4.5C shows estimates of grand means of recollection and familiarity for each bias condition. Clearly, neither recollection nor familiarity estimates differ across the biased-out and biased-in conditions.
Figure 4.5: Results from Experiment 2. A. ROC curves for the biased-in and biased-out conditions, constructed by averaging over participant and items. B. Estimates of DPSD criteria for the biased-in condition as a function of those for the biased-out condition for each participant. Points labeled “1” are the criteria separating the “sure new” and “believe new” responses; points labeled “2” are criteria separating the “believe studied” and “sure studied” responses. C. Estimates of the grand means of recollection and familiarity \( (d^{(s)} - d^{(n)}) \) for the biased-in and biased-out conditions. Error bars are 95% credible intervals.

### 4.2.3 Discussion

Using payoffs to manipulate response biases for certain confidence ratings over others appears to move points along a common ROC curve. Moreover, the payoff manipulation selectively influenced the criteria parameters in the DPSD model. Beyond being an example of selective influence in DPSD, these results provide novel evidence that confidence ratings are tapping the same response biases as other means of drawing ROC curves (e.g., differential payoffs).

Experiment 2 provides evidence that criteria may be selectively influence in DPSD. The goal of Experiments 3, 4, 5, & 6 is to determine whether recollection and familiarity parameters may also be selectively influenced, as claimed by Yonelinas (2002). If instead changes in recollection are always accompanied by changes in familiarity, this redundancy would suggest that a single-parameter model (e.g., non-normal signal detection) is sufficient to describe the data.
4.3  Experiment 3: LOP & Match-Mismatch

According to the dual-process model, a levels-of-processing (LOP) manipulation (Craik & Lockhart, 1972) at study should affect only recollection. Alternatively, a perceptual match-mismatch manipulation should affect only familiarity. These predictions were tested in Experiment 3 by varying both LOP and perceptual match-mismatch conditions within subjects and fitting the hierarchical DPSD model.

4.3.1  Methods

Participants

Fifty-two University of Missouri students participated in Experiment 3 in return for credit toward a course requirement.

Stimuli

Stimuli were the same 480 words used in Experiment 1.

Procedure

The procedure was identical to Experiment 1 with the following exceptions. At study, each word was accompanied by one of two sets of instructions. For half of the words, the instructions read “Read the word out loud and enter the number of vowels”. For these shallow study instructions, participants used the numeric keyboard keys to enter in the number of vowels in the word on the screen. For the remaining words, the instructions read “Read the word out loud and enter a related word”. For these deep study instructions, participants were instructed to enter in the first word that came to mind which was related to the word on the screen. Instructions varied randomly from trial to trial. Each study word was either displayed in lowercase letters and red
ink, or in uppercase letters in green ink. Following study, the test phase consisted of
240 new words and the 240 previously-studied words. The color/case of half of the
studied words was the same as when the word was studied, and the color/case of the
remaining half was different (e.g., study in lowercase-red, tested in uppercase-green).
This match vs. mismatch condition was counterbalanced with LOP condition.

4.3.2 Results

ROC curves for each condition (averaged over participants and items) are shown in
Figure 4.6A. A crude measure of accuracy may be constructed by calculating hits as
any “studied” response to studied items, and false alarms as any “studied” response
to new items. The difference of the inverse probit transforms of these rates (averaged
over items) provides a crude estimate of $d'$ (from EVSD). Doing so reveals that overall
accuracy for two participants was near chance ($d' < .1$), and so their data were not
considered in analysis. The average $d'$ values for the remaining 50 participants were
subjected to a 2x2 mixed-effects ANOVA with LOP condition and match-mismatch
condition serving as factors. Although the effect of match-mismatch was small, the
ANOVA revealed main effects of both match-mismatch ($F(1, 49) = 9.7, p < .05$) and
LOP ($F(1, 49) = 405, p < .05$), but no interaction between the two ($F(1, 49) = 0.74$).

The hierarchical dual-process model was fit to the data with the R package HBMEM.
Figure 4.6B shows estimates of recollection for each of the four conditions as a function
of estimates of familiarity $d'$. The large effect of LOP is primarily in recollection,
as predicted by the dual-process model. Familiarity estimates for the shallow and
deep LOP conditions are near 1.0, and so the lack of any substantial effect of LOP
on familiarity does not reflect a floor or ceiling effect. Unfortunately, the match-
mismatch effect was very small, as can be seen in the ROC curves (Figure 4.6A).
In the DPSD model fit, this small effect manifests only in familiarity, as predicted.
However, because the effect is so small it is possible that a lack of any match-mismatch
effect in recollection is merely a symptom of having a small effect to begin with. In Experiments 4-6, I attempt to increase the size of the match-mismatch effect in order to determine whether it really does provide evidence of selective influence.

The same items were used in Experiments 1 and 3. If parameters of the hierarchical DPSD model are measuring meaningful aspects of the data, then item-effect estimates should be similar across these experiments, even though they employ different participants and different study conditions (recall that in Experiment 1 participants simply read each word aloud). In Figure 4.7, item effect estimates from Experiment 3 are plotted as a function of those from Experiment 1. Clearly, there is a relationship between these estimates for new-item familiarity, studied-item familiarity, and to a lesser extent, recollection. This comparison does not provide strong support for the DPSD model — if another model such as gamma signal-detection was generating the data, we would nevertheless expect DPSD parameters to be similar across experiments. Figure 4.7 does, however, bolster our claim that item variability is systematic and consistent, and therefore, ought to be accounted for in any ROC analysis.
Figure 4.7: DPSD item effects from Experiment 3 plotted as a function of those from Experiment 1. A. Item effects on new-item familiarity \(d_j^{(n)}\). B. Item effects on studied-item familiarity \(d_j^{(s)}\). C. Item effects on recollection \(R_j\).

The left panels of Figure 4.8 show recollection effects for participants (top) and items (bottom) as a function of sensitivity \(d'\). As in Experiment 1, recollection is positively correlated with sensitivity for both participants and items. The middle and right panels show the relationship between recollection and new-item familiarity, and recollection and studied-item familiarity, respectively. Again, as in Experiment 1, the relationship between recollection and familiarity is primarily driven by a negative correlation between recollection and baseline familiarity for both participants and items. As discussed in Chapter 3, this relationship may reflect structure underlying the two memory processes. Alternatively, it may reflect the fact that a single process is generating these data, and fitting an overly complex model manifests as correlations among parameters.

In addition to fitting the DPSD model, the gamma, UVSD, and extended UVSD signal detection models were fit to the data from Experiment 3. The dual-process model outperformed UVSD with one value of \(\sigma^2\) by 534 DIC points, the gamma model by 497 points, and the extended UVSD model with free \(\sigma^2\) across people and items by 465 DIC points. These patterns in DIC fit are nearly identical to those in Experiment 1; the magnitudes of the DPSD advantage is smaller simply because there are fewer participants.
Figure 4.8: The relationship between recollection and familiarity estimates from Experiment 3 for participant effects (top) and item effects (bottom). The left panels show recollection as a function of sensitivity $d'$ defined as $d^{(s)} - d^{(n)}$. The middle panels show recollection as a function of new-item, baseline familiarity $d^{(n)}$, the right panels show recollection as a function of studied-item familiarity $d^{(s)}$.

4.3.3 Discussion

The results of Experiment 3 show that the levels-of-processing effect is primarily in the recollection parameter, as predicted by the dual-process model. There is also, however, a small LOP effect in familiarity, possibly indicating a 1-process model that manifests as large effects in recollection accompanied by small effects in familiarity (in this range of the space). Unfortunately, the match-mismatch effect was too small to be useful in assessing selective influence. For example, it clear that in Figure 4.6B a line can be drawn through the recollection-familiarity space that accounts for all four conditions in Experiment 3. It is not clear, however, whether this uni-dimensional structure could be violated had a larger match-mismatch effect been observed. The
goal of Experiments 4a-c is to increase the magnitude of the match-mismatch effect so that selective influence may be better assessed.

4.4 Experiments 4a-c: Increasing the Match-Mismatch Effect

In the Yonelinas (2002) review, the match-mismatch effect provides the most compelling evidence of a manipulation that affects familiarity without affecting recollection (see Figure 4.1). A closer look at the data used in this meta-analysis reveals that 1) all of the studies utilized the remember-know procedure, and 2) three of the four studies came from Gregg and Gardiner (1994). Moreover, in the fourth study, performed by Rajaram (1993), there was exactly zero effect of match-mismatch. In Experiments 4a - c, I attempt to more closely replicate the work of Gregg and Gardiner (1994) in an effort to increase the match-mismatch effect. In Experiment 4a, all words were studied visually on the computer screen (as in previous experiments). At test, however, some words were presented visually as they were during study, and some were presented auditorily over headphones. Studied words are in the perceptual match condition when they are tested visually, and are in the mismatch condition when tested auditorily. Our hope is that this modality (visual vs. auditory system) match-mismatch manipulation will produce larger effects than the case-color manipulation used in Experiment 3.

4.4.1 Experiment 4a: Modality Match-Mismatch

Methods

Twenty-six University of Missouri students participated in Experiment 4a in return for credit toward a course requirement. Stimuli were the same 480 words used in
Experiment 1. In addition to the visual stimuli, each word was recorded being read by a male voice as a 2-second audio file. At study, participants viewed words visually and performed one of the two LOP instructions used in Experiment 3 (count vowels vs. generate related word); participants were not instructed to read each word aloud. All of the words were presented in white on a black background. At study, half of the words were presented on the computer screen, and participants made one of six confidence ratings. The remaining half of words were presented auditorily over a pair of headphones. One second prior to, and during auditory presentation, the screen read “Audio Presentation” accompanied by a picture of headphones. Following the 2-second auditory presentation, “Respond To Audio Presentation” appeared on the screen, at which point participants made a confidence rating.

**Results**

Data for two participants were discarded due to a computer error. ROC curves, constructed by averaging data over participants and items, are shown in Figure 4.9A. A crude measure of $d'$ was calculated as it was in Experiment 3. A within-subjects ANOVA on $d'$ with LOP and match-mismatch conditions as factors revealed a main effect of LOP condition ($F(1, 23) = 358.60, p < .05$), a marginally significant main effect of match-mismatch condition ($F(1, 23) = 4.14, p = .054$), and no interaction ($F(1, 23) = 2.83, p = .11$). Values of $d'$ for the match and mismatch conditions (averaged over LOP condition) were 2.02 and 1.87, respectively. The goal of Experiment 4a was to increase the size of the match-mismatch effect compared with that of Experiment 3, but as the ANOVA and ROC curves in Figure 4.9A shows, this goal was clearly not met. Thus, no further analyses were conducted.
4.4.2 Experiment 4b: Fast Study Presentation

Experiment 4a was an attempt to make Experiment 3 more like the work of Gregg & Gardiner’s first experiment, in which they claimed to have found an effect of modality match-mismatch. However, the effect was also small in their first experiment, and they increased the effect in their second experiment by making the study task a “highly perceptual orienting task”. At study, words were presented on the screen for 200ms, and participants were asked to simply count how many of these words appeared blurry (none of the words were actually blurry). The goal was to make participant’s encoding of the words highly perceptual, thus increasing the modality match-mismatch effect. We followed Gregg & Gardner in our Experiment 4b, in another attempt to increase the modality match-mismatch effect.

Methods

Six University of Missouri students participated in Experiment 4b in return for credit toward a course requirement. We had intended to run 25 participants, but the data from these six made it clear that the match-mismatch effect was no larger than in Experiment 4a. Stimuli were the same 480 words used in Experiment 1. At study, participants were shown 280 words for 200 ms each, with a 300 ms blank period
between words. They were instructed to simply look at each word, and press the space bar during presentation of the word if they were unable to read it. Following study, participants were briefed that their memory would now be tested for the words that were presented. The test procedure was identical to that of Experiment 4a (half of the words presented visually, half presented auditorily). Note that participants were not briefed about the test phase until after study, making this an incidental learning task.

Results

ROC curves, constructed by averaging data over participants and items, are shown in Figure 4.9B. Crude estimates of $d'$ were computed for each participant as they were in Experiment 4a. The difference in performance between the match ($d' = .42$) and mismatch ($d' = .44$) conditions was not significant ($t(5) = .50, p = .64$). Although the sample size is small, a Bayes factor t-test (see Rouder et al., 2009) on these $d'$ values reveals that they are 2.9 times more likely to come from the same distribution than different ones. Clearly, there was no modality match-mismatch effect, even though the study task was designed to be highly perceptual. This lack of effect, however, may be due to the overall poor performance in this task. If performance is near floor for both conditions, then there would be no opportunity for an effect to arise even if one exists.

4.4.3 Experiment 4c: Shorter Study-Test Lists

Experiments 4a and 4b differed from those of Gregg & Gardiner in that our study lists contained 240 items with 480 items at test; they used 40 and 80 item study and test lists, respectively, across several study-test sessions. It is possible that longer test lists force participants to rely on semantic memory rather than perceptual, negating any
modality match-mismatch effect. Moreover, the longer study and test lists explain the poor accuracy obtained in Experiment 4b. Experiment 4c was the same as 4b, except that participants performed several smaller study-tests sessions, rather than one big session.

Methods

Sixteen University of Missouri students participated in Experiment 4c in return for credit toward a course requirement. The experiment was identical to Experiment 4b with the following exception. Participants studied 40 words, and were then tested on 80. They completed this study-test cycle 6 times.

Results

ROC curves are shown in Figure 4.9C. Average performance for the match ($d' = 1.13$) and mismatch ($d' = 1.02$) conditions was higher than in Experiment 4b, however, they did not differ significantly ($t(15) = 1.76, p = .10$). Again, although there is a small modality match-mismatch effect, it is much too small to be of use for assessing selective influence.

4.4.4 Discussion

The meta-analysis of Yonelinas (2002) implies that perceptual match-mismatch manipulations affect familiarity without affecting recollection. Such an effect would provide especially good evidence for the dual-process model, as finding an effect in familiarity alone is rare or even non-existent (see Rouder, Lu, et al., 2008). Unfortunately, a closer look this meta-analysis reveals that all of the experiments that actually showed an effect at all were from Gregg and Gardiner (1994), who used the remember-know paradigm. The goal of Experiments 4a-c was to replicate Gregg &
Gardiner in the confidence rating paradigm. Unfortunately, across three experiments, either no effect of modality match-mismatch was found, or the effect was much too small for making reliable inferences about selective influence.

In a more recent study, Boldini, Russo, and Avons (2004) examined the modality match-mismatch effect under conditions of speeded and un-speeded responses in an old-new recognition memory task. When participants were allowed ample time to make old-new judgments, Boldini et al. found no effect of match-mismatch, as in Experiments 4a-c above. However, when participants were forced to make old-new judgments within a short time window (less than 700 ms), the modality match-mismatch effect was as large as .5 $d'$ units. Because recollection takes time to develop, theoretically, forcing participants to respond quickly eliminates any contribution of recollection to recognition judgments. Because participants must rely solely on familiarity to perform the task, manipulations that affect familiarity (such as modality match-mismatch) become apparent.

In Experiment 5 I attempt to replicate the speeded conditions of Boldini et al., in which a match-mismatch effect was found, in the confidence-rating paradigm. This experiment provides three important tests of the dual-process model. First, the effect of speeding responses should be in recollection and not familiarity. This prediction will be assessed by comparing the results of Experiment 5 with those from Experiment 3, in which responses were not speeded. The second prediction is that, because recollection has been reduced or eliminated, there should be no effect of LOP in the speeded conditions, as LOP effects should be mediated by recollection alone. Finally, Experiment 5 should serve as a replication of Boldini et al., but with confidence ratings rather than simple old-new recognition. Specifically, there should be a match-mismatch effect even though none was found in Experiment 3 or Experiments 4a-c.
4.5  Experiment 5: Response Deadline; Long Study List

Experiment 5 was identical to Experiment 3 in that there was both an LOP manipulation and a perceptual match-mismatch manipulation. It differed from Experiment 3 in that participants were required to make confidence-rating responses within a short window following presentation of each test word.

4.5.1  Methods

Participants

Fifty-three University of Missouri students participated in Experiment 5 in return for credit toward a course requirement.

Stimuli

Stimuli were the same 480 words used in Experiment 3, presented visually throughout the experiment.

Procedure

The study session in Experiment 5 was identical to that used in Experiment 3. Participants either counted the number of vowels in each study word (shallow LOP condition), or generated a word that was related to each study word (deep LOP condition). Each study word was also presented in lowercase-red, or uppercase-blue. At test, half of the studied words were presented in the same case/color as at study (perceptual match) or the opposite case/color (perceptual mismatch). Each test trial began with a fixation cross. Following 500 ms of fixation, a test word appeared. After 500 ms, "****" was presented below the test word. Participants were instructed to
respond as soon as possible, after the “****” appeared. If the response was made
before the “****” appeared, a low-pitched error tone (buzz) sounded, followed by the
message “Too Fast, Respond after ****”, presented for 3 seconds. If the response
was made more than 500 ms after the “****” appeared, the same low-pitched error
tone sounded, followed by the message “Too Slow, Response Faster!!!”, also presented
for 3 seconds. Thus, participants had to respond within a window from 500-1000 ms
after the test word was presented. For comparison, in Experiment 3 the average re-
sponse time was 1.97 seconds, and the 25th percentile was 1.05 seconds. Thus, the
500-1000ms window should allow ample time to make a confidence rating response,
but not allow for much deliberate reflection upon whether a word had been studied
or not.

Prior to the study phase, participants completed two training phases to get them
acquainted with the response window and test instructions. First, participants were
given test trials with “WORD” serving as the stimulus. They were instructed to
make any of the 6 confidence ratings as quickly as possible after the “****” appeared.
They had to make their response within the correct response window 15 times in a
row before moving on (most participants had no problem doing so after making a few
“too slow” errors). Next they completed a practice study-test phase with a 10-item
test list, before moving on to the study phase.

4.5.2 Results

Responses that were made outside of the response window (13%) were discarded. A
crude measure of accuracy was constructed as for previous experiments. Doing so
reveals that overall accuracy for five participants was near or below chance ($d' < .01$)
indicating a failure to follow instructions, and so their data were not considered in
analysis. ROC curves for each condition (averaged over participants and items) are
shown in Figure 4.10A. The crude $d'$ values for each participant were subject to a 2x2
mixed-effects ANOVA with LOP condition and match-mismatch condition serving as factors. As in Experiment 3, although the effect of match-mismatch was small, the ANOVA revealed main effects of both match-mismatch ($F(1, 47) = 5.69, p < .05$) and LOP ($F(1, 47) = 109.96, p < .05$), but no interaction between the two ($F(1, 47) = 0$).

To determine the effect of response speeding on the match-mismatch effect, the match-mismatch effects (as measured with the binomial d’ value) may be compared across the un-speeded (Experiment 3) and speeded (Experiment 5) conditions. Comparing the mean match-mismatch effect for participants in Experiment 3 (.10) with those from Experiment 5 (.06) reveals that there is no significant difference between them ($t(96) = 1.02, p = .31$).

The hierarchical dual-process model was fit to the data with the R package HBMEM. Figure 4.10B shows estimates of recollection for each of the four conditions as a function of estimates of familiarity $d'$. Because according to the dual-process model speeding should affect mostly recollection, and the LOP effect should be in recollection, there should be no effect of LOP in the speeded conditions. Figure 4.10 shows that
there is some effect of LOP, however, the magnitude of this effect is greatly attenuated when compared with the un-speeded conditions of Experiment 3. The remaining effect of LOP appears to be in both recollection and familiarity, however, it is not clear that either of these differences are significant. The small match-mismatch effect again appears to manifest only in familiarity, as predicted by the dual-process model. As in Experiment 3, however, this effect is too small to make reliable claims about selective influence. It is, however, interesting that in four of the four conditions tested, the small match-mismatch effect is seen only in familiarity.

Figure 4.11 shows estimates of recollection as a function of familiarity for Experiments 1, 3 (un-speeded), and 5 (speeded) collapsed over match-mismatch condition. Whereas the dual-process model predicts that the effect of speeding should be primarily in recollection, speeding seems to affect both recollection and familiarity. It is, however, possible that the response window was so short as to not only degrade the recollection process, but perhaps also did not allow participants enough time to make responses based on familiarity. The critical question is whether all of the explored effects (LOP, match-mismatch, and response speeding) manifest in both recollection and familiarity, such that a line may be drawn through the 2-dimensional space reducing it to one dimension. If so, then this reduction is evidence that the dual-process model is over-specified. Across these three experiments, a monotonically-increasing line may be drawn that encompasses each of the nine conditions. One such line is shown in Figure 4.11. Although this simple line is not completely encompassed by all credible intervals, it is close, and a more complex line exists that could do so and yet retain monotonicity.

### 4.5.3 Discussion

The results of Experiments 1, 3, and 5 do not provide evidence in favor of the dual-process model. The effects of the LOP manipulation, and speeded vs. un-speeded
response deadline conditions were both large. Both effects, however, were present in both recollection and familiarity. The result is that recollection and familiarity parameter estimates trace out a 1-dimensional curve in Figure 4.11 that is characteristic of a single-process model.

The line in Figure 4.11 implies a single-process model. It is possible, however, that conditions exist with parameters that lie off the common 1-dimensional line, but I have simply not found them yet. Experiment 6 is an effort to find such a condition. Looking at Figure 4.11 it is clear that the existence of any condition that has low familiarity but high recollection would violate the current 1-dimensional nature of the space. In ROC space, such a condition would be a nearly straight line that lies
substantially off of the diagonal. Although to my knowledge such data have only been observed in a working memory task (Rouder, Morey, et al., 2008), one plausible way to achieve such a condition is to increase the amount of recollection in Experiment 5 without affecting familiarity. Although not shown in the Yonelinas (2002) review, one purported way to selectively affect recollection is to vary the length of the study (and test) list (Yonelinas, 1994). Increasing the number of studied items increases inter-item interference. According to the dual-process model, such interference should only affect recollective processes, as familiarity is automatic and operates independently of other storage processes (Yonelinas & Jacoby, 1994). In Experiment 6, Experiment 5 is replicated but with a shorter study list. The dual-process model predicts that this effect should be solely in recollection. If such selective influence in achieved, then these conditions will lie in the upper-left part of Figure 4.11, thus violating the uni-dimensionality of the space.

4.6 Experiment 6: Response Deadline; Short Study List

The goal of Experiment 6 is to replicate Experiment 5, but with a shorter study list. According to the dual-process model, study list length should affect only recollection. If this predicted selective influence holds, then the recollection and familiarity estimates will lie outside of the 1-dimensional line in Figure 4.11.

4.6.1 Methods

Participants

Sixty-four University of Missouri students participated in Experiment 6 in return for credit toward a course requirement.
Stimuli

Stimuli were 160 words randomly selected from the 480-word pool used in previous experiments.

Procedure

Experiment 6 was identical to Experiment 5 with the exception that each participant studied 80 words (selected randomly) and was tested on 160. The study and test lengths were therefore one third the length of those used in Experiment 5.

4.6.2 Results

Responses that were made outside of the response window (15%) were discarded. A crude measure of accuracy was constructed as for previous experiments. Doing so reveals that overall accuracy for eight participants was near or below chance ($d' < .01$) indicating a failure to follow instructions, and so their data were not considered in analysis. ROC curves for each condition (averaged over participants and items) are shown in Figure 4.12A. The $d'$ values for each participant were subject to a 2x2 mixed-effects ANOVA with LOP condition and match-mismatch condition serving as factors. As in Experiments 3 & 5, although the effect of match-mismatch was small, the ANOVA revealed main effects of both match-mismatch ($F(1, 55) = 5.23$, $p < .05$) and LOP ($F(1, 55) = 113.3$, $p < .05$), but no interaction between the two ($F(1, 55) = .52$, $p = .47$).

Figure 4.12B shows estimates of recollection for each of the four conditions in Experiment 6 as a function of estimates of familiarity $d'$. Like Experiment 5, the effect of LOP appears to be in both recollection and familiarity. Unlike previous experiments, the match-mismatch effect also now appears in both recollection and familiarity. This effect, however, is still too small to allow for any serious conclusions.
Figure 4.12: Results from Experiment 6. A. ROC curves for each condition. B. Estimates of recollection for each condition plotted as a function of estimates of familiarity \( (d^s_k - d^n_k) \) for each condition.

regarding its locus.

Figure 4.13 shows familiarity and recollection estimates from Experiments 1, 3, 5, and 6 (collapsed over match-mismatch conditions). There are two critical features in the figure. First, comparing estimates from Experiment 5 with those from Experiment 6 reveals that the list-length effect was only in the recollection parameter, as predicted by the dual-process model. Second, the points from Experiment 6 do not lie on the line that was able to be drawn through the data from Experiments 1, 3, and 5. There may exist a monotonically increasing line that can intersect the points from all four experiments shown in Figure 4.13. This line, however, would require very sharp curves (i.e., approaching non-monotonicity), and even then would just barely include all of the 95% credible intervals.

The condition from Experiment 6 in which encoding was deep, participants were forced to respond quickly, and the list length was short violates the pattern seen in all other conditions. The importance of this single point, then, warrants close inspection. Figure 4.14A shows estimates of recollection as a function of baseline strength for each
Figure 4.13: DPSD estimates from Experiments 1, 3, 5, & 6. Estimates of recollection for each condition are plotted as a function of estimates of familiarity, averaged across match-mismatch conditions. The line is same function shown in Figure 4.11.

participant in this condition; Figure 4.14B shows recollection as a function of studied-item familiarity. The relationship between recollection and baseline familiarity is similar to that observed in Experiments 2 and 4. There is also, however, a negative relationship between recollection and studied-item familiarity, whereas no relationship between these estimates was observed in previous experiments. This relationship is driven by about a dozen outliers, shown in gray in Figure 4.14B. The critical question is what makes these participants different than all other participants observed thus far.

Figure 4.14C shows response distributions for the outlying participants in Figure 4.14B, and Figure 4.14D shows distributions for non-outliers. Clearly, whereas
most participants distributed responses across the confidence ratings, the outliers use the “sure new” and “sure studied” responses almost exclusively. These data are extremely problematic for model fitting — they have only two degrees of freedom, and so the ROC curve has only two points. An ROC with only two points can be perfectly accounted for by either of two models. First, an equal-variance signal-detection model, which forms the familiarity component of the dual-process model, can account for these data perfectly. Second, a simple all-or-none high-threshold model, by which participants respond “sure studied” if they detect that an item was studied, and guess “sure studied” or “sure new” if they don’t, can account for the data perfectly. As a result, the dual-process model can account for these data perfectly in two ways: 1) by having no recollection and high familiarity (pure EVSD), or 2) by having high recollection and no familiarity (high-threshold).

With two ways to perfectly account for the data (high recollection and no familiarity, or vice versa), the model must choose one. If the DPSD model were being fit separately to participants, then this choice would be a coin flip for each. In the hierarchical model, however, participants who have extreme parameter values are pulled in toward the group average for that parameter. In Experiment 6 the average value of recollection across all participants is rather high, and so it is assumed that the outlying participants also have high values of recollection. Accordingly, a model by which these participants have no recollection but high familiarity is deemed implausible. Instead, the hierarchical model allows these participants to have low values of $d_i^{(s)}$, and assumes that they must have high values of recollection. Given the data from all other participants, it is simply more plausible that these outliers have high recollection and no familiarity than vice versa. In reality, however, the data for these outliers give us no information regarding their levels of recollection or familiarity, and so should not be considered as coming from the same group as the other participants.

The method for determining which participants are outliers is not obvious; perhaps
one “believe” response is enough for the DPSD model to provide reasonable estimates; perhaps several are necessary. To error on the side of caution, I considered participants outliers if they responded “sure” 90% or more of the time. This criterion results in 23 outlier participants in Experiment 6, 14 in Experiment 5, 5 in Experiment 3, and 2 in
Figure 4.15: Estimates of recollection & familiarity from model fits to data excluding participants who only used the “sure” responses. A. Points are the same as in Figure 4.13, but without outlying participants. B. Same as A, but data are not averaged over match-mismatch condition, and recollection estimates are drawn on probit scale ($\Phi^{-1}(R)$). The line is the best-fitting unweighted least-squares regression line.

Experiment 1. The hierarchical DPSD model was fit to the data from the remaining participants in each of these four experiments. Estimates of recollection as a function of familiarity from these analyses are shown in Figure 4.15A. Clearly, considering only those participants who’s data are rich enough for model fitting, the results across three experiments, with four conditions in each, are perfectly in line with a single-process explanation of recognition memory. Figure 4.15B shows these estimates for every condition (i.e., not averaged over match-mismatch), with recollection plotted on the probit scale. It appears that the underlying relationship between recollection and familiarity may be linear in this space.

4.7 Discussion

The goal in Chapter 4 was to determine whether parameters of the dual-process signal detection model could be selectively influenced. In Experiment 2, it was shown that
a response-bias manipulation indeed affects only criteria parameters. Although the
effect of a levels-of-processing manipulation in Experiments 3, 5, and 6 manifested
mostly as an effect in recollection (as predicted), there was also evidence for a smaller
but consistent effect of LOP in familiarity. Speeding responses (Experiments 5 &
6) had a large effect on recollection (as predicted), but again had a sizable effect on
familiarity estimates as well. In Experiments 3, 4a-c, 5, and 6, the effect of perceptual
match-mismatch was exceedingly small or non-existent, even though several of these
experiments were specifically designed to increase the effect. Although it has been
claimed that the effect of perceptual match-mismatch affects only familiarity (Gregg
& Gardiner, 1994; Yonelinas, 2002), we failed repeatedly to find a sizable enough
effect to warrant such a claim. Moreover, other researchers have also failed to find
a match-mismatch effect (Boldini et al., 2004; Mulligan et al., 2010), and it is likely
that the few cases where an effect is found represent a selection bias to publish non-
null effects. If such small effects are partitioned into different processes, such as
recollection and familiarity, the resulting minuscule effects in each parameter can not
be used to claim selective influence. Thus, our results cast serious doubt on the claim
that perceptual match-mismatch selectively influences familiarity. Accordingly, there
exists no convincing demonstration of a manipulation affecting familiarity and not
recollection, or vice versa.

The effects of LOP, match-mismatch, and response speed all manifested as effects
in both recollection and familiarity. At first glance, however, the effect of list length
under a response deadline appeared to be only in recollection. Upon closer inspection,
in this condition several participants adopted a response strategy whereby they used
only the outermost response options. If such data are plotted in ROC space, the curve
would consist of only two points. Consequently, these data are perfectly explained
with either a pure EVSD model, or a pure high-threshold model. When these data
are fit with the dual-process model, the prior structure on recollection and familiarity

128
determines whether these participants have all recollection and no familiarity, or all familiarity and no recollection. Because such estimates certainly do not reflect the processes of interest, data from these participants should not be considered in analysis. When the dual-process model is fit to data for participants who used the inner confidence ratings at least 10% of the time, then every manipulation considered affected both recollection and familiarity. Moreover, the recollection familiarity space is clearly reducible to one dimension (see Figure 4.15).
Chapter 5

Summary and Concluding Remarks

The goal of this dissertation was to explore the processes underlying recognition memory using hierarchical Bayesian models. In this final chapter I review the extent to which this goal was met.

In Chapter 1 I showed how averaging data over people or items that vary in their mnemonic processes underlying recognition memory judgments could distort conclusions. In particular, if items that vary in $d'$ are averaged over, then data from an equal-variance signal detection model will produce asymmetric ROC curves. This asymmetry would be confounded with the mnemonic process, for example, by interpreting the amount of item variability (asymmetry) as a measure of explicit recollection.

In Chapter 2 I introduced a hierarchical unequal-variance signal detection model. This UVSD model explicitly accounts for both participant and item variability, thus circumventing any need to average data over participants or items. Application of this model to a large confidence-rating experiment revealed that ROC asymmetry remained even after partialling out variability due to participants and items. This result was incongruent with the prediction from Chapter 1 because items and participants do not simply vary in $d'$, but rather, exhibit nearly the same amount of
variability in baseline strength and studied strength. This result, and several others, could not have been obtained without the hierarchical model. Moreover, even though asymmetry does reflect mnemonic process, averaging data over participants or items still resulted in incorrect estimates, and confidence intervals on those estimates that were biased too small. The result is that analyses utilizing averaged data will provide far too much confidence in distorted results.

In Chapter 3, I developed an extended version of the hierarchical UVSD model that allowed for participant and item variability in $\sigma^2$. Fitting this model to Experiment 1 revealed that $\sigma^2$ does indeed vary across participants and items. This variability, however, was systematic such that increases in sensitivity $d'$ were correlated with increases in ROC asymmetry $\sigma^2$. This structure was accounted for with a simple gamma signal detection model that, by assuming mnemonic strengths were distributed as gamma distributions that differed in scale, predicted a relationship between sensitivity and asymmetry without the need for any more parameters than EVSD. A hierarchical version of the Yonelinas dual-process model was also developed which allowed participant and item variability in both recollection and familiarity. Fitting this model to Experiment 1 revealed similar patterns as the UVSD fit — recollection and familiarity were positively correlated across people and items. The abilities of these models to account for Experiment 1, taking parsimony into account, were compared via DIC. This analysis provided strong support for the dual-process model over all single-process competitors considered.

In Chapter 4 I assessed whether parameters of the dual-process model could be selectively influenced. In addition to relying on item-averaged data, previous studies claiming selective influence are wrought with methodological flaws. In Experiment 2, manipulations that were intended to affect participants’ response biases for certain confidence ratings over others properly affected DPSD criteria, but not recollection or familiarity. Six additional experiments were conducted in an effort to affect rec-
ollection and familiarity independently. Across these six experiments, the effect of perceptual match-mismatch was too small to be useful for making any claims about selective influence. Thus, it seems that there is no manipulation that affects familiarity without affecting recollection. In Experiments 3, 5, and 6, a levels-of-processing manipulation was found to affect both recollection and familiarity, although in each case there was more evidence for an effect in recollection than familiarity, as predicted. Comparing experiments in which participants were forced to make confidence judgments within a small response window with experiments that had no time constraints revealed that the effect of response-deadline was in both recollection and familiarity. Finally, under speeded conditions, the effect of study list length was also in both recollection and familiarity. Thus, considering manipulations of match-mismatch, levels-of-processing, response deadline, and list length, neither recollection nor familiarity were able to be influenced in isolation.

In addition to assessing selective influence, the results from Chapter 4 allowed for an assessment of the dimensionality underlying recognition memory. If a single process is generating the data, then estimates of recollection and familiarity should be correlated across all constructs. In particular, a plot of recollection as a function of familiarity should be able to be reduced to one dimension by drawing a monotonically-increasing line through all data points. Figure 4.15A shows recollection estimates as a function of familiarity estimates from Experiments 1, 3, 5, and 6. Clearly, the two dimensions of recollection and familiarity may be reduced to one. Moreover, Figure 4.15B shows that the relationship between familiarity and the probit transform of recollection may be a simple linear function.

The dual-process model provided a superior fit to the data, taking model complexity into account, in two experiments (Exp. 1 & 3) according to the DIC statistic. However, fits of the DPSD model clearly suggest that the data are generated from a one-process model. There are two possible reasons for this discrepancy. First, the
DIC statistic performs well when comparing two models that have similar numbers of parameters, such as DPSD and UVSD with a different $\sigma^2$ for each person and item. DIC is a hierarchical extension of AIC, and like AIC, it is biased to toward more complex models. It may therefore not be well-suited for comparing models that differ greatly in their complexity, such as the DPSD and gamma signal detection models. The second possibility is that there exists a single-process model that provides a better account of the data than the DPSD model, but that is was not considered it in the model comparisons in Chapter 3. Interesting candidates include the gamma signal detection model with a shape other than 2.0, and other non-normal signal detection models such as the Gumbel (DeCarlo, 1998). Alternatively, non-signal detection accounts such as mixture models (Broder & Schutz, 2009) may prove promising.

In conclusion, all of the data collected herein are perfectly compatible with a single-process interpretation of recognition memory, including the effects of items, people, levels-of-processing, perceptual match-mismatch, and slow vs. speeded responding. There is therefore no convincing evidence that recognition memory is mediated by two processes. It is my hope that until such evidence is presented, this model will cease being used to interpret data from both normal and clinical populations. These are important pursuits, but the current trend is to construct theories based on fits of the dual-process model. If this model is wrong, as a close inspection of the data suggests it is, then the field of memory will make little progress until it is abandoned.


Michael Sylvester Pratte was born in Saint Louis Missouri on September 21, 1981, the son of Yvonne and Gary Pratte. After completing high school at DeSmet (St. Louis) and Blue Valley (Kansas City), he attended Kansas University. He completed his Bachelor of Arts degree in Psychology, with a minor in Philosophy, at the University of Missouri in Columbia. He went on to complete his Masters and PhD in Psychology at the University of Missouri with Jeffrey Rouder as his adviser. During this time he was married to Sarah Garber, and together they raised Olive the English Setter. Following defense of his dissertation, Michael will begin a Postdoctoral position with Frank Tong at Vanderbilt University.