March 2019

Testing Recognition Memory Models with Forced-choice Testing

Qiuli Ma

TESTING RECOGNITION MEMORY MODELS WITH FORCED-CHOICE TESTING

A Thesis Presented

by

QIULI MA

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

February 2019

Psychology

TESTING RECOGNITION MEMORY MODELS WITH FORCED-CHOICE TESTING

A Thesis Presented

by

QIULI MA

Approved as to style and content by:

_____
Jeffrey J. Starns, Chair

_____
Caren Rotello, Member

_____
Rebecca Ready, Member

_____
Caren Rotello, Chair of the Psychological and Brain Sciences Department

ABSTRACT

TESTING RECOGNITION MEMORY MODELS WITH FORCED-CHOICE TESTING

FEBRUARY 2019

QIULI MA, B.Eng., XI'AN JIAOTONG UNIVERSITY

B.A., XI'AN JIAOTONG UNIVERSITY

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jeffrey J. Starns

People's ability to call an experienced item "old" and a novel item "new" is recognition memory. Recognition memory is usually studied by first asking participants to learn a list of words and then make judgments of old (studied) or new (not studied) for test words. It has long been debated whether the underlying process of recognition memory is continuous or discrete. Two types of models are compared specifically that assume either discrete or continuous information states: the 2-high threshold (2HT) model and the unequal variance signal detection (UVSD) model, respectively. Researchers have used the receiver operation characteristic (ROC) function and response time (RT) data to test between the two models. However, both methods have provided evidence for 2HT and UVSD, and the debate has not come to consensus. In this study, we used an alternative approach to look into this issue. After studying the words, participants first made "old/new" judgment for each single test item. Then, if there were falsely identified items, each of them was randomly paired with a correctly identified word of the same response. Participants were asked to choose the studied word from the word pair. Simulation and experimental results were able to discriminate the 2HT and UVSD model. Experimental results

showed that the UVSD model fitted the data better than the 2HT model. The forced-choice test paradigm provided an effective way to test between the 2HT and UVSD models.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE CONTINUOUS AND DISCRETE ACCOUNTS OF RECOGNITION MEMORY

1.1 Introduction

Recognition memory is a form of declarative memory. It concerns people's ability to tell whether something is "old" – meaning that they have experienced it before in a specified context – or "new" – meaning that they have not. Recognition memory is usually tested with word lists. In an experiment, participants first study a list of words. Later another set of words is shown to them. Participants respond "old" or "studied" if they think the tested word was on the study list; they respond "new" or "not studied" if they think the word was not on the list. Studied words are called targets; not studied words are called lures.

Recognition memory has been extensively studied with mathematical modeling (e.g., Pazzaglia, Dube, & Rotello, 2013; Kellen, Klauer, & Bröder, 2013; Starns, Ratcliff, & McKoon, 2012). One fundamental question in the modeling literature concerns the nature of the information retrieved from memory. There are two influential modeling approaches that starkly disagree on this question: multinomial processing tree (MPT) models and signal detection theory (SDT) models. MPT holds that recognition decisions are informed by several discrete inner cognitive states (e.g., Swets, 1961; Snodgrass, & Corwin, 1988; Luce, 1963), whereas SDT holds that those decisions are the result of comparisons between decision criteria and memory strength values drawn from a continuous distribution (Green, & Swets, 1966).

The most successful form of MPT model has been the two high-threshold (2HT) model (Snodgrass, & Corwin, 1988). According to this model, a target word can lead to two mental states. If any evidence showing the word is on the study list is remembered, the detect old mental

state is reached which leads to an "old"/"studied" response. If no evidence is remembered, the guess state is reached. For a lure word, if it provides any information proving its absence on the list, a detect new mental state is reached which leads to a "new"/"not studied" response. If no discounting information is retrieved, then the guess state is reached. Once target or lure test words enter into the guess state, they will be treated equally. Since no evidence is recollected about them, "old"/"studied" or "new"/"not studied" responses will be made by pure guessing. The critical property that defines the model as a high-threshold process is that targets never enter into the detect new state, and lures never enter into the detect old state.

In the 2HT model, for a given target, a subject enters the "detect old" state with probability d_o, yielding an "old" response. For a given lure, the subject enters the "detect new" state with probability d_n, yielding a "new" response. With probability $1 - d_o$ and $1 – d_n$, the subject enters a state of uncertainty and guesses "old" with probability g and "new" with probability 1- g. So the detection (*d*) parameters represent how effectively participants remember the items, and the guessing (*g*) parameter represents response biases.

In contrast with the 2HT model, a standard signal-detection model assumes a continuous distribution of memory evidence for both targets and lures (Green, & Swets, 1966; Macmillian, & Creelman, 2005). As shown in Figure 1, the mean of target distribution is greater than that of the lure distribution, reflecting the fact that memory evidence tends to be stronger for these items. A decision criterion is set along the dimension of memory strength, and is usually denoted with λ. Recognition decisions are made based on the retrieved strength value's relative location to λ: if the evidence falls on the right of the criterion, an "old" response is made, and if it falls on the left, a "new" response is made. A basic version of the SDT recognition memory model includes two equal-variance Gaussian distributions of memory strength across items, i.e., the EVSD model.

However, in practice, an unequal variance signal detection (UVSD) model is found to be a better account quantitatively (Wixted, 2007). In UVSD, the target distribution's standard deviation is larger than the lure's, indicating that more variability is added in when subjects have gone through extra phases of studying.

The means of the target and lure distribution are denoted by μ_t and μ_l, and their variances σ_t and σ_l. The mean and standard deviation of the lure distribution are set to 0 and 1 by convention. The criterion λ can move along the dimension ranging from the most liberal at the left end to the most conservative at the right end.

The 2HT model has been used by vast majority of recognition memory studies in the MPT field. Other models, such as the low threshold (LT) account, have only been recently considered. Also, the UVSD model outperforms EVSD for its better ability to fit the recognition data (Egan, 1958; Ratcliff, Sheu, & Gronlund, 1992; Wixted, 2007). So in our study, the 2HT model from MPT family and the UVSD model from the SDT family will be compared.

1.2 Testing the Models with ROC Functions

The 2HT and UVSD models have been primarily tested in terms of their ability to match the receiver operating characteristics (ROCs) (see Yonelinas, & Park, 2007, for a review). An ROC function plots the hit rate (the proportion of "old" responses among targets) against the false alarm rate (the proportion of "old" responses among lures) across different levels of response bias or confidence, reflecting the subject's willingness to say "old" to test items.

The 2HT model produces linear ROC functions, where the hit rate is $d_o + (1- d_o)g$ and false alarm rate is $(1 – d_n)g$. The line intersects the *y* and *x* axes at $(0, d_o)$ and $(1 – d_n, 1)$,

respectively. Predicted hit and false-alarm rates move up along this line as the probability of guessing "old" increases.

For the UVSD model, the hit rate is $1 - p(\lambda, \mu_t, \sigma_t)$ and false alarm rate is $1 - p(\lambda)$. *p()* is the cumulative probability density function of a Gaussian distribution. μ and σ are the mean and variance of memory evidence for targets, and λ is the decision criterion. When the criterion decreases, the area of the target and distractor distributions that falls to the right of the criterion (and thus the proportion of "old" responses) both increase. As a result, both hit rate and false alarm rate increase. However, due to the nonlinearity of the Gaussian distribution, the hit rate and false alarm rate do not increase with the same rate, making the ROC function a curve rather than line. In the ROC function, this is reflected as a convex curvature between (0, 0) and (1, 1). Furthermore, because the standard deviations of target and lure are different, this curvature is asymmetrical.

Previous studies have largely shown support for the ROC of recognition memory to be curved (Yonelinas, & Park, 2007). This is because subjects are able to retrieve some information about the test item even when they respond at low confidence levels. This is made possible by the overlapping nature of the target and lure distributions in SDT, which is contrary to 2HT model's assumption that low-confidence responses are the result of guesses in the absence of clear evidence identifying a target or a lure.

Malmberg (2002) pointed out that 2HT can also generate curved confidence rating ROCs because the slope of two adjacent points on the ROC function was determined by the ratio of the probability of a target being assigned to a certain confidence level and the probability of a lure being assigned to that level. Those probabilities could change across confidence levels. If the mapping of "detect old" and "detect new" responses were allowed to vary across confidence

levels instead of all responses being set at the highest confidence level, the ratio would vary across different levels, thus producing a curved function. Consequently, the conclusion that curvilinear ROCs supported the continuous model rather than discrete state model was not necessarily valid for confidence rating ROCs.

More recently researchers have focused on the bias manipulation ROCs. Bröder and Schutz (2009) reanalyzed 59 data sets in the literature that manipulated response bias in recognition via payoffs or base rates in recognition experiments. For data sets with two-step bias manipulation, they assumed equal variances in the SDT model and equal detect old and new rate in the 2HT model. For data sets with more than two bias levels, they fit the UVSD model and the 2HT model with different detect old and detect new parameters. They found no apparent advantage for the SDT or 2HT model, so the 59 data sets did not reject 2HT in favor of SDT. Bröder and Schutz also conducted three recognition experiments with 5-step bias manipulations and kept the encoding and testing conditions equivalent. They fitted the data with seven-parameter versions of both models, including sensitivity, standard deviation and the five bias steps. They compared the G^2 statistics and found that the 2HT model was able to fit the data better than the SDT models. They argued that the SDT and 2HT models were equally valid as measurement tools for recognition memory.

A reanalysis of Bröder and Schutz's study by Dube and Rotello (2012) found that among the 62 cases in Bröder and Schutz's meta-analysis, 43 of them varied biases over only two levels. As two-point ROCs could be fitted with either a curve or a line, misfits of both models were unable to provide distinguishable information. Moreover, among the remaining 19 cases that had 3 to 5 bias levels, 15 supported the SDT models.

Dube and Rotello (2012) also pointed out that in one of their three new experiments, where Bröder and Schutz showed that the SDT model was outperformed by 2HT, instead of using word stimuli as in the first experiment, they used line drawings, whose different coding operation could usually produce more linear ROCs (Onyper, Zhang, & Howard, 2010).

Dube and Rotello then reported two newly designed experiments that closely examined the confidence rating ROCs and bias manipulation ROCs. A large number of trials were collected which enabled comparisons of the two models on both individual and aggregated ROCs. The goodness-of-fit indicators supported the UVSD model.

In order not to solely rely on the goodness-of-fit parameters that lack the consideration of different models' flexibilities in model fitting, other indices were introduced. Among them were the Akaike information criterion (AIC) (Akaike, 1973; Wagenmakers, & Farrell, 2004) and the Bayesian information criterion (BIC) (Schwarz, 1978; Anderson, & Burnham, 2002). However, as the two indices both determined model flexibilities based on the number of free parameters, the 2HT and UVSD model would be treated as equally flexible for many studies, even though the models might differ in their true flexibility to match noise in the data. Therefore, some researchers have turned to a more comprehensive measurement called normalized maximum likelihood (NML). NML contains two components. The first component corresponds to the maximum log-likelihood of the observed data in a particular experiment, representing model fit. The second one is a penalty factor that is the sum of the maximum log-likelihoods of all possible data patterns that could be observed from the experiment. These two terms correspond to the two considerations in model fitting: overfitting and generation error respectively, trying to find out the best fitting model that is also the most parsimonious one. Kellen et al. (2013) reanalyzed 41

datasets and multiple models of recognition memory with NML, and concluded that models from the 2HT family were most supported by the individual level analyses.

1.3 Non-ROC Tests of the Models

Another important aspect of recognition memory decision is the response time (RT), which has been used to test between 2HT and UVSD models in a few studies. The diffusion model assumes a continuous evidence accumulating process over time, rather than several discrete inner cognitive states for recognition (Ratcliff, 1978). This is consistent with the UVSD model. Small steps of evidences are accumulated towards two boundaries that correspond to "old" or "new" responses. The distance between two boundaries reflects the speed-accuracy trade-off. The starting point of the accumulator is affected by manipulations of bias. There have not been too many studies that focused on the RT account of the 2HT model, but it would be reasonable to conjecture from the model that the more links on a MPT tree a test item has to go through, the larger RT it will take (Hu, 2001). Taking the two models together, it can be seen that manipulating biases will affect both the shape of ROC functions and RT distributions. Implementing this idea, Dube and Rotello (2012) conducted two experiments with bias manipulations and found out that UVSD was supported over 2HT with both ROC fitting and response time modeling.

Another prediction that the 2HT model makes for the RT data is that study-strength should only affect the detect probability of a studied item, but not its mapping to different response states nor response times once the information state is determined. This is called the conditional independence assumption (Province, & Rouder, 2012). In contrast, the SDT model would predict that strength values farther from the criterion would result in faster RTs. Province

and Rouder (2012) used a two-alternative forced-choice (2AFC) paradigm tested this prediction. In a typical forced-choice study, participants first view a list of words and then complete the memory test. In contrast to single-word recognition, two words were presented side by side in each trial of forced-choice. One word was studied and the other was not. Participants' task was to indicate which one of the two was the target. Providence and Rouder found that the conditional mean RT, which was the RT for targets that entered the detect state, did not vary with the number of study opportunities. This supported the 2HT model. Another study tested this theory with both group and individual RTs (Kellen, Singmann, Vogt, & Klauer, 2015). The mean RTs replicated Province and Rouder's result (2012). The individual RTs, which were tested with a linear mixed-model (LMM) where the conditional probability that a response was produced by a certain state was used as the covariate, showed significant effects on RT of conditional detection-probability when study-strength was controlled but not the other way around. In other words, the results were consistent with the conditional independence in that RT was not predicted by the encoding condition after the internal state produced by the item was statistically controlled.

1.4 Forced-choice study

Several studies have used ROC functions from a forced-choice task to test alternative models (Province, & Rouder, 2012; Kellen, Singmann, Vogt, & Klauer, 2015; Jang, Wixted, & Huber, 2009; Kellen, & Klauer, 2011; Kroll, Yonelinas , Dobbins, & Frederick, 2002; Parks, & Yonelinas, 2009; Smith, & Duncan, 2004). For instance, in Jang et al., (2009), after studying a single list of words, participants were tested with yes/no word recognition trials and 2AFC test trials. Participants responded to both forms of test with 6-level confidence rating. Apart from the UVSD model, two additional models, DPSD (dual-process signal detection) and MSD (mixture

signal detection) were compared. The DPSD model contains a threshold-like, high confidence recollection process and a continuous familiarity process that equals to an EVSD model. The mixture model also assumes a continuous value of memory strength, but its target distribution is a mixture of two equal variance Gaussian distributions with different means. The three models were simultaneously fit to recognition and 2AFC data for each participant with parameter constraints derived from the 2AFC and recognition test response relationship. The UVSD model was found to be the best model among the three to describe the relationship between yes/no and 2AFC recognition performance. The UVSD model also provided the best fit to participants' performance considering model flexibility. The DPSD model outperformed the MSD model.

Parks, & Yonelinas (2009) and Kellen, & Klauer (2011) used four-alternative forced-choice task with two responses (4AFC-2R) to distinguish recognition memory models by examining the accuracy of the second response when the first response was a lure. In the test, participants were presented with four words and alternatively tested on "standard" 4AFC trials and second choice trials. On the "standard" 4AFC trials, participants chose one word out of four as the target word. On the second choice trials, they made two ordered responses, where the first response was the most likely to be the studied word, and the second response was the next most likely to be the studied word. There were three categories of response: first choice incorrect and second choice correct, first choice incorrect and second choice incorrect, and first choice correct and second choice incorrect. The UVSD model provided better fit to the response patterns than the other models. Model fitting results favored the UVSD model over EVSD, DPSD, MSD and threshold models, with model complexity analyses of the NML method.

Kellen and Klauer (2014) also conducted a ranking study. Participants completed a four-alternative ranking task and a three-alternative task in two experiments. In this task, participants

rank the test items according to their belief that they were studied. For example, participants may be presented with three words on a screen, they are asked to assign 1, 2, and 3 to each word, with 1 representing the word is mostly to be on the study list, and 3 representing the word is least likely to be on the study list among the three. In Kellen and Klauer (2014), there were two types of word stimuli presented randomly during the test. Weak stimuli were words studied once, and strong stimuli were words studied three times. The SDT model predicted that the conditional probability of a studied item being assigned to the second rank given it was not assigned to the first rank increased with memory strength. The 2HT model predicted that this probability stayed constant as item strength changes. Kellen and Klauer stated that the ranking judgment provided an alternative comparison method between memory models with several advantages. It did not require model fitting and parameter estimation. There was also no need for distributional assumptions, exhaustive experimental manipulations and complex model selection methods. The experimental results were found to be more consistent with SDT model's prediction.

As previous research did not come to consistent conclusions, there needs to be some novel methods to distinguished the two models. Our study also used the forced-choice test to discriminate recognition memory models, but with critical differences. In previous research, a stimulus was either tested in the single-item recognition test, or the forced-choice test. There is not much connection between the single-item recognition and the forced-choice test. The two are essentially separate tests. In our study, the single-item and forced-choice tests made use of the same set of stimuli. Each forced-choice trial was consisted of a target and a distractor that had the same response during the single-item recognition. So one word was correctly recognized and the other one was incorrectly recognized. As detailed below, this procedure allowed us to test

specific predictions about forced-choice accuracy conditional on the outcome of the earlier single-item trials.

As the 2HT and UVSD models assume different mechanisms of recognition error, determining which model is more consistent with the real performance data can help us better understand why recognition errors happen. The 2HT model assumes that recognition errors happen because of unlucky guesses when people fail to remember anything from past experience. The UVSD model assumes that recognition errors happen because of misleading information retrieved from memory. These two reasons suggested by the 2HT and UVSD respectively are totally different from each other. Moreover, recognition memory models are useful measurement devices that can help us answer research questions, such as whether different populations (young and older adults) differ in their memory abilities. Different models can make different conclusions about research questions, so it is important to determine which models make assumptions that are most consistent with observed data. Knowing which model is true not only allow people understand more about the psychological reason of their behavior, but could also help them make less recognition errors in real life.

CHAPTER 2

EXPERIMENT

2.1 Introduction

We conducted two experiments that tested the UVSD and 2HT models without using ROC functions. After learning a list of words, participants first completed the single-item recognition task; that is, they saw a single word appear on the computer screen and decided whether it was "old" (seen on the study list) or "new" (not seen before). In the first experiment, participants responded to this word without bias and in the second experiment, participants were encouraged to respond with conservative or liberal biases. Biases were manipulated with payoffs. In the conservative condition, participants gained 1 point or lost 3 points for correct and incorrect "old" responses, respectively, and gained 3 points or lost 1 point for correct and incorrect "new" responses, respectively. In the liberal condition, participants gained 3 points or lost 1 point for correct and incorrect "old" responses, respectively, and gained 1 point or lost 3 points for correct and incorrect "new" responses, respectively. The purpose of the bias manipulation was to test contrasting qualitative predictions of the two models, as explained below. Participants' responses at different levels of bias reflected their willingness to call a test word "old" (or "new").

After the single-item test, participants were brought to a forced-choice phase where they were shown a target and a lure, one of which was previously classified incorrectly. Some trials were "old"-"old" (O-O) trials, comprising a target that was correctly called "old" and a lure that was incorrectly called "old." Some trials were "new"-"new" (N-N) trials, comprising a target that was incorrectly called "new" and a lure that was correctly called "new." So the condition labels (O-O and N-N) refer to the previous response that the participant made for both items, not

the actual items in the trial (which was always one target and one lure). Participants were asked to choose which word was studied. The UVSD and 2HT models make different predictions for the forced-choice data, as described in the next section.

2.2 Model predictions

Take the situation that both words are called "old" in the single-item test as an example (i.e., an O-O forced-choice trial). According to the 2HT model, the probability that a target is called old in single-item recognition is $p(old) = d_o + (1 - d_o)g$; that is, the probability that it will be detected as "old" plus the probability that the detection will fail but the participant will guess "old." Therefore, the proportion of targets called "old" that were detected as old is $p = \frac{d_o}{d_o + (1 - d_o)g}$. On the other hand, if a lure is called old, it has to be a guessing error, indicating no evidence is retrieved about this item. When the two words called "old" are presented together in a forced-choice trial, the probability that the target will be picked as the "old" one is $p + (1 - p) * 0.5$. In other words, on p trials participants will select the target because they detected that it is old, and on the remaining trials they have to randomly choose an item, leaving them with a .5 chance of selecting the target.[1] Substituting p into the previous equation reveals that the forced-choice percent correct is a linear function of the guessing parameter g: $0.5 + \frac{0.5d_o}{d_o + (1 - d_o)g}$. With $d_o < 1$, the percent correct of forced choices decisions goes down as guessing parameter g increases in the O-O task. Similarly, in a N-N task (both target and lure

[1] We assume that the information available from memory for an item is the same when it is tested in the single-item trials and when it is tested in the forced-choice trials. In the experiments, there will be a short lag between two test trials with the same item, so it is unlikely that the memory state will change drastically.

have been called "new" in the single-item test), the probability that a lure is selected as "new" is also a linear function of g: $0.5+\frac{0.5d_n}{1+(d_n-1)g}$. With $d_n < 1$, the percent correct goes up as g increases. This trend is illustrated in Figure 2. The forced-choice test percent correct plotted against the probability of guessing "new" (1-g); thus, values farther to the right reflect more conservative single-trial responding. As responding becomes more conservative, the percent correct of O-O trials increases and N-N trials decreases.

The predictions stated above are also psychologically plausible: if people are more willing to guess new, then large portion of "old" responses to targets would have come from the detect-old state. When shown together with a lure that has been called "old" in an O-O forced-choice trial, participants will most likely select the target word as "old" because it is usually a word that they detected was on the list. However, if people are less willing to guess new, there will be more guesses among "old" responses, including the "old" responses of targets. Because all stimuli in the guessing state have no memory evidence retrieved, when such a target is shown together with a lure that has been called "old", participants would have to make a random guess again. In short, the forced-choice percent correct increases when participants are more willing to guess new (conservative), and decreases when participants are less willing to guess new (liberal). For N-N task, lures called "new" will be more likely to be based on detection when the participant is less likely to guess "new". Thus, participants will be more likely to recognize the lure and respond correctly in N-N tasks. Contrary to the O-O task, the forced-choice percent correct of N-N task is higher when participants are less willing to guess "new" (liberal) than when they are more willing to guess "new" (conservative).

In the UVSD model, a lure is called "old" because its evidence strength falls to the right of the decision criterion. In order for the target to be correctly picked out in an O-O forced-

choice trial, the target's evidence strength has to be even further to the right of the criterion than its paired lure. So in an old-old task, the probability that a target is selected as old is $\int_{\lambda}^{\infty} \frac{d(x,\mu_t,\sigma_t)}{1-p(\lambda,\mu_t,\sigma_t)} \frac{p(x)-p(\lambda)}{1-p(\lambda)}$. In this equation, $d()$ indicates a probability density functions and $p()$ indicates a cumulative probability density function of a Gaussian distribution. For example, $d(x,\mu_t,\sigma_t)$ is the likelihood of strength value (x) on the target distribution given its mean (μ_t) and standard deviation (σ_t). The first fraction is the likelihood that a *recognized* target has strength value x, and it is found by dividing all likelihoods above the recognition criterion by the probability that a target is called "old" (meaning that it is above the criterion). The second fraction is the probability that a *falsely recognized* lure has a strength value below x, and it is found by dividing the proportion of all lures between the criterion and x by the proportion of all lures above the criterion. In other words, $p(x)-p(\lambda)$ represents all possible lures that have been falsely recognized as old but have smaller memory strength than the target. $1 - p(\lambda)$ represents all lures that have been falsely recognized. Multiplying the two fractions gives the joint probability that a recognized target has strength value x and has a higher strength value than a randomly selected lure that was falsely recognized. Integrating over x gives the total probability that any recognized target would have a higher strength value than a lure called "old," corresponding to the predicted forced-choice percent correct. Likewise, in a new-new task, the probability that the lure is recognized is $\int_{-\infty}^{\lambda} \frac{d(x,\mu_t,\sigma_t)}{p(\lambda,\mu_t,\sigma_t)} \frac{p(x)}{p(\lambda)}$.

It is hard to get a simple linear relationship between the forced-choice percent correct and the decision criterion λ. However, it can be inferred that when the decision criterion λ increases, lures called "old" in the old-old task will have higher strength. Under many parameterizations, this makes it more likely the lure's strength will exceed that of the target's, so the lure is more

likely to be incorrectly selected as "old". In other words, percent correct decreases for old-old trials as single-item responding gets more conservative for most parameter values. In the new-new task, when decision criterion increases, targets called "new" can come from regions with higher memory strength. When randomly paired with a correctly rejected lure, it is more likely that this high memory strength will exceed the strength of the lure. In other words, the percent correct of the new-new forced-choice task will increase as criterion goes up. This prediction holds for all parameter sets, so it is more general than the old-old prediction. An illustration of the change is plotted in Figure 3, where the target distribution has a mean of 1 and standard deviation of 1.2.

We simulated 20,000 sets of randomly selected model parameters to make predictions of forced-choice data for the 2HT and UVSD models. The parameters were sampled from uniform distributions in the following way: Mean of the target distribution (μ_t) of UVSD model varied from .4 to 1.8; standard deviation varied from 1.1 to 1.4. To calculate the bias criteria, first a random value was generated from the uniform distribution between $\frac{\mu}{2} - .25$ and $\frac{\mu}{2} + .25$ to serve as the halfway point between the criteria. Then a distance was randomly sampled between .2 and .8. The conservative criterion was calculated by adding half the distance to the halfway point, and the liberal criterion was calculated by subtracting half the distance from the halfway point. Detect old and new probabilities of the 2HT model were both allowed to vary in uniform distributions between .25 and .6. To calculate the two guessing parameters, first a value was randomly sampled between .3 and .7 to serve as the halfway point. A distance was drawn between .2 and .6. The conservative and liberal guessing parameters were calculated by subtracting and adding half the distance to the halfway point, respectively.

Figure 4 shows percent correct difference between liberal and conservative bias trials. The first panel shows that in the 2HT model, conservative trials' forced-choice percent correct was always greater than that of the liberal trials' for O-O trials. But for N-N trials, the forced-choice percent correct was always greater in liberal trials than in conservative trials. The opposite was true in the UVSD model. In O-O trials the liberal trials' percent correct was mostly greater than that of the conservative trials, and in N-N trials, it was conservative trials that always had the greater percent correct. Figure 4 suggests that not only the way that the percent correct changes as the bias level changes was opposite for O-O and N-N forced-choice trials, it was also opposite for the 2HT and UVSD models.

The above showed that different models predicted different patterns of the forced-choice data. Thus in this study we used a forced-choice task to determine whether the 2HT or UVSD is a better recognition memory model. Figure 4 showed that data from our forced-choice test could discriminate the two models without model fitting or using sophisticated statistics for fitting evaluations. The critical data we analyzed in the forced-choice test was straightforward and easy to compare. Thus, our paradigm is a simple but very powerful way to differentiate the two models of 2HT and UVSD.

2.3 Method

2.3.1 Participants

In the simulations described below, the proportion of simulated participants that were better fitted by the true model than the alternative model was .83 for 2HT and .73 for UVSD in Experiment 1, and .90 for 2HT and .83 for UVSD in Experiment 2. Taking the lowest value of .73, in order to have over 90% power for binomial tests to determine if one model fits better to

more participants than would be expected by chance (with alpha = .05), at least 40 participants' data are required.

A total of 88 UMass Amherst undergraduate students participated in Experiment 1. We recruited more participants than needed because the experiment data would also be used for a study involving RT analysis. We kept all participants' data to achieve high power for this study. The participants were recruited through the SONA system and received experimental credit in exchange for their participation.

Experiment 2 had 45 participants. Among them 39 were recruited from the SONA system and 6 from an online advertisement posted on the psychology department website. Participants attended two sessions of Experiment 2. SONA participants received 1 credit per session they attended. The other 6 participants received $12 for each session they attended.

2.3.2 Materials

Words for the study and test lists were randomly sampled from 1098 nouns from the SUBTLEXus dataset (Brysbaert & New, 2009). The word frequency ranges from 10.02 to 99.49 per million words based on subtitles from American films and television series. Each participant completed three study-test cycles per session, with new words for each cycle. They were encouraged to respond as accurately as they could.

There were two types of trials in the test phase: single-item trials and forced-choice trials. In a single-item trial, a word showed up on the computer screen for participants to decide if it was on the study list or not. Participants hit "z" key to respond "new" and "/" key to respond "old," and they were asked to keep their index fingers on the "z" and "/" key throughout the test.

After the single-item test, words that were incorrectly classified were paired with correct ones to make forced-choice trials. For example, a lure that was falsely recognized as "old" would be randomly paired with a correctly recognized target, i.e., a target that was responded with "old." Participants were asked to indicate which one of the two was studied. Table 1 shows an example of a cycle of single-item and forced-choice trials.

There are 3 incorrect trials: "restaurant," "spectacular," and "disaster" in Table 1. In this case, "restaurant" was randomly paired with a correctly responded target, (e.g., "way") to make an O-O forced-choice trial. The other two incorrectly classified targets ("spectacular" and "disaster") were each paired with a word randomly selected from "lady," "flight," or "obedient" to make N-N force-choice trials. Participants chose one word from the two that they thought was indeed studied.

If there were not enough correctly responded targets to make O-O forced-choice pairs, the number of such trials was determined by the number of correct target trials. The same was true for the N-N test. So the number of forced-choice trials was determined by both the number of incorrect trials and their corresponding correct trials, whichever was smaller.

2.3.3 Procedure

2.3.3.1 Experiment 1

In Experiment 1, participants went through three study-test cycles. The first cycle was practice where participants studied 28 words. The real cycles each had 68 studied words, among which 8 words were fillers. Studied words were grouped into blocks of four. So a total of 68 studied words appeared in 17 blocks. Fillers appeared in the first two blocks. We included fillers in study list to control for primacy effects.

In study phase, each block of four words was followed by a recall task that prompted participants to type in a word from the block with the keyboard. The word was identified by its order in the block, and every word had equal probability of being probed. The purpose of the recall task was to make sure that participants focused on learning the words throughout the study phase. None of the words probed for recall were used as targets on the subsequent recognition test to avoid increasing variability in memory across items given that recall should improve subsequent memory. In this case, discarding one word from every block, every block contained 3 "real" target words. Subtracting words from the two filler blocks, a study list ended up contributing 45 target words towards the following recognition test.

Immediately following the last study block, a screen appeared prompting participants to begin the single-item test. A word showed up on the computer screen with two choices: "not studied" and "studied" respectively. Participants hit the "z" key to respond "new" and the "/" key to respond "old."

Because words recalled during the study phase were not tested in recognition, there were 90 trials in single-item recognition with 45 targets and 45 lures (94 if one includes the four filler trials that began the test – 2 targets from the filler study blocks and 2 filler lures). After every 10 real trials of single-item test, participants were prompted to start the forced-choice test. If it was the first time for the participant to run through the forced-choice test, she or he would read a short instruction about the forced-choice test. Participants were informed that all words in the forced-choice test came from the past 10 items. To make sure that participants read both words before they responded to the forced-choice word pair, the two words first appeared one at a time for 1000 ms each on the computer screen, during which time participants did not have chance to respond until the two words appeared simultaneously side by side with response alternatives on

the screen. In the forced-choice test trials, words that were incorrectly classified during the single-item test were randomly paired with ones that were correctly classified with the same response category. Participants hit the "z" key if the studied word was on the left, or hit the "/" key if it was on the right. After the forced-choice trials were complete, participants were prompted to start the next test block by subsequently hitting the "z" and "/" key, respectively.

2.3.3.2 Experiment 2

Participants completed 2 sessions on different days for Experiment 2. Each session consisted 3 study-test cycles. The first cycle of the first session was a practice cycle. When studying, the first and last 4 words on list were always fillers. So there were 8 fillers for every study list. The study list of the practice cycle contained 40 words, and the real cycles contained 72 words (including 8 fillers). Unlike Experiment 1, there was no recall test in study phase.

During the single-item test trials, participants responded with bias towards or against "studied" response in two types of blocks. Bias was manipulated by payoffs. In liberal blocks, participants gained 3 points for correct "old" responses and 1 point for correct "new" responses; they lost 1 point for incorrect "old" responses but lost 3 points for incorrect "new" responses. In conservative blocks, participants gained 1 point for correct "old" responses and 3 points for correct "new" responses; they lost 3 points for incorrect "old" responses and 1 point for incorrect "new" responses. Responses that earned participants 3 points if they were correct and lost 1 point if they were incorrect were labeled as safe responses in the instructions, and were indicated with green font throughout the test phase. Responses that earned participants 1 point if they were correct but lost 3 points if they were incorrect were labeled as risky responses, and were indicated with red font. For example, on conservative blocks "Not Studied" appeared in green

and “Studied” appeared in red to label the alternative responses on each trial. Before every block, participants were notified whether the “Studied” or “Not Studied” response was the safe response in the following block.

To make sure that participants attend to the bias information during the single-item test, they were asked to make two responses. First, “studied” and “not studied” choices appeared on the screen before the test word was presented. Participants made a response solely based on the bias information. It was anticipated that the green colored response would be chosen more often since it was the safe one. After the first response, the test word was presented with the “old/new” choices. Participants made the second response based on their memory about the word and bias information. When participants made more than one risky response guess and lost points, they saw feedback on how many points they had lost due to irrational guessing, and they were informed that the better strategy was to always guess the safe (green) response than the risky (red) response. This was again to make sure that participants respond according to the bias information.

In single-item recognition, participants went through 48 “real” test trials and 4 filler trials in practice cycles. They went through 4 fillers and 96 “real” trials in real cycles. The first 4 test words were always fillers. After every 12 real trials, the forced-choice test immediately followed if there were incorrectly recognized words. Each study-test cycle contained 8 single-item and forced-choice test blocks. Experiment 2 terminated when participants had earned at least 1,286 points in total or they had been in the experiment for 45 minutes. 1,286 points was a high standard that equaled to a situation when participants’ every single-item test response and half of the guessing responses were correct. So participants were not able to leave the experiment until they had completed all three study-test cycles. Any additional trials after the third cycle were not

analyzed. Participants completed 2 sessions of Experiment 2 on different days. The first cycle on the first day was a practice cycle.

2.4 Simulations

We conducted simulations with exact the same conditions and trial numbers as the actual experiments to explore the predictions of the models. Data were generated from the 2HT and UVSD models, and then each model was fitted to the simulated data. G^2 was used as the indicator of goodness-of-fit. If the 2HT model fits the performance better, its G^2 would be smaller than that of the UVSD model. The opposite would be true if UVSD fits the result better.

2.4.1 Parameters

In the UVSD model, the lure distribution was conventionally set as a normal distribution with a mean of 0 and a standard deviation of 1. To sample simulated data sets, the mean of the target distribution (μ_t) was allowed to vary between .4 and 1.8 and the standard deviation (sd_t) was allowed to vary between 1.1 and 1.4. For Experiment 1 where we did not manipulate the response bias, the single criterion (λ) was randomly sampled from a uniform distribution between $\frac{\mu}{2} - .25$ and $\frac{\mu}{2} + .25$. For Experiment 2 that has two bias conditions, the distance between the criteria (the size of the criterion shift) was drawn from a uniform distribution between .2 and .8 and the halfway point between the criteria was drawn from the same distribution used for λ in Experiment 1. The conservative criterion was calculated by adding half the distance to the halfway point, and the liberal criterion was calculated by subtracting half the distance from the halfway point.

To generate simulated data sets from the 2HT model, d_o and d_n were sampled from the uniform distributions between .25 and .6. For Experiment 1, the guessing parameter was drawn from uniform distribution between .3 and .7. Again, for Experiment 2 with two bias conditions, a difference in the guessing parameter between conservative and liberal blocks was drawn from the uniform distribution of .2 and .6 and the halfway point between them was sampled from the same distribution used for the guessing parameter in Experiment 1. The conservative and liberal guessing parameters of biases were calculated by subtracting and adding half the distance to the halfway point.

2.4.2 Simulation results of Experiment 1

Data were simulated for 2,000 participants. Each simulated participant went through 180 trials of single-item test (90 targets and 90 lures), the same number of trials as the empirical subjects in Experiment 1.

Both the 2HT and UVSD models were used to fit the data. When simulation data were generated by the 2HT model, the summed G^2 was 8,212 for the UVSD model, and 2,199 for the 2HT model across all participants. In a head-to-head comparison of the UVSD and 2HT models, 1,663 (83%) out of 2,000 participants were better fit by the 2HT model. When the simulation data were generated by the UVSD model, the summed G^2 was 2,113 for the UVSD model, and 5,428 for the 2HT model. Among the 2,000 participants, 1,451 (73%) were better fit by the UVSD model.

In this simulation, the G^2 of all simulated participants should follow a χ^2 distribution with 1 degree of freedom (4 response frequencies minus 3 free parameters) if the model fitted to the data was the true model (plotted with red lines in the first two panels of Figure 5 and 6). Figure 5

shows the model fitting results when the data were generated with a UVSD model. The first two panels show histograms of G^2 when the data were fitted with the UVSD and 2HT models, respectively. The third panel plots the G^2 differences between the two fits. It can be seen that the spread of G^2 was smaller for the UVSD fits than for the 2HT's, and the median of the G^2 differences (indicated with a red line in the third panel) between the two was less than 0, meaning that the UVSD model has better fits than 2HT model.

Figure 6 shows the results when the data were generated with a 2HT model. The first two panels also show the G^2 histograms when the data were fitted with the two models respectively. The data were better accounted by the 2HT model, as reflected by the smaller G^2 values in the 2HT fits 2HT panel. The median of G^2 differences between the UVSD and 2HT fittings was greater than 0, showing that 2HT fitted the data better than UVSD.

Simulation results in Figure 5 and 6 show that the performances of the two models, 2HT and UVSD, were separable in Experiment 1. When the participants' data are fitted with the two models, the fitting results will be more like Figure 5 if the UVSD model better captures the data pattern, which will suggest that UVSD model describes the underlying recognition memory mechanism better than the 2HT. If the 2HT model better captures the data, the fitting results will be more like what's presented in Figure 6. This demonstrates that evaluating the relationship between single-item and forced-choice performance provides a new way to test the models that is independent of bias ROC tests. With our design, the models are discriminated even without a bias manipulation.

2.4.3 Simulation results of Experiment 2

To keep consistent with the real Experiment 2, 192 trials of single-item test (96 targets and 96 lures) were simulated with 2,000 participants.

When the simulation data were generated by the 2HT model, the summed G^2 was 19,787 when fitted by the UVSD model, and 8,310 when fitted by the 2HT model. In a head-to-head comparison of the UVSD and 2HT models, 1,804 (90%) out of 2,000 participants were better fit by the 2HT model. When simulation data were generated by the UVSD model, the summed G^2 was 8,509 when fitted by the UVSD model, and 17,129 when fitted by the 2HT model. Among the 2,000 participants, 1,663 (83%) were better fit by the UVSD model.

In Experiment 2, the G^2 should follow a χ^2 distribution with 4 degrees of freedom (8 response frequencies minus 4 free parameters) when the model fitted to data was the true model. Histograms of the G^2 values give more intuitive comparison. Figure 7 shows the model fitting results when the data were generated with a UVSD model. The first two panels are G^2 histograms when the data were fitted with the UVSD and 2HT model respectively. Like in Figure 5, the G^2 was smaller when the data were fitted by UVSD model than when it was fitted by the 2HT model. Again, the median of G^2 differences was less than 0, indicating the UVSD models were able to fit the data better than the 2HT models. Figure 8 is similar to Figure 6, where the data were generated by the 2HT model and was better fitted by the 2HT models.

Simulations of Experiment 2 showed that with the 2-step bias manipulation, the performances of the UVSD and 2HT models were distinguishable. If the UVSD model better accounts the experimental data, G^2 values for the UVSD model fits will be smaller than that of the 2HT model; if the 2HT model better accounts the data, G^2 values for the 2HT model fits will be smaller than that of the UVSD model.

2.5 Results

2.5.1 Experiment 1

Of all the 88 UMass students who participated in Experiment 1, 19 participants data were dropped because their percent correct of single-item recognition was below .6. Participants' performances are reported in Table 2. The mean percent correct of the single-item recognition is .69. The mean percent correct of the forced-choice trials was .64. The fact that the percent correct of forced-choice test was smaller than single-item recognition might seem surprising at the first sight, given that forced-choice testing provides the benefit of getting information from two items instead of one. However, our forced-choice trials were constructed in such way that both stimuli had the same previous responses. This was very likely to have made the task more difficult. Parameters of the best fit model across all participants were reported in Table 3.

2HT and UVSD models were used to fit the data. The summed G^2 across all participants was 98 for the UVSD model and 244 for the 2HT model. The summed G^2 should follow a χ^2 distribution with 69 degrees of freedom, assuming the fitted model was true. Both models' summed G^2 value went past .95 quantile of this comparison χ^2distribution (89); thus, the fits are unexpectedly bad if one assumes that the fitted model is the true model producing the empirical data. This suggests that the two models both failed to fit some participants' data. Notably, the 2HT model missed much more participants than the UVSD. Among the 69 participants, the UVSD model produced a better fit than the 2HT model for 55 participants (80%, $p < .001$ by binomial test). For individual G^2s, 10 (14%) participants' values were significant with χ^2 test ($p = 0.05$) with the 2HT model, meaning 14% of the participants were not fit well by the assumed model. For UVSD model, 2 (3%) participants' values were significant ($p = 0.05$). Since the

proportion of misfitting data, 3%, was even less than the α level (5%) we assumed, the UVSD model seems to be describing the data well at the individual-participant level.

G^2 histograms are shown in Figure 9. The fitting results of the UVSD and 2HT model were plotted in the first and second panel respectively. The χ^2 distribution that G^2 should follow for a true model is indicated with red lines. The third panel plots the G^2 differences between the two fits. It can be seen that the spread of G^2 was smaller for UVSD fits than for 2HT's, and the median of the G^2 differences (indicated with a red line in the third panel) between the two is less than 0. It suggests that the UVSD model fits the data better than 2HT does, in line with the result of the binomial test.

It is possible that participants' memory might have changed during the period between the single-item test and the forced-choice test. In that case, the assumption that participants make forced-choice judgment based on the same evidence retrieved during single-item test would not be valid. This could produce a spurious fitting advantage for the UVSD model over the 2HT model, for reasons I will now explain.

According to the 2HT model, the forced-choice task can be easier than the single-item recognition. In comparing the two stimuli that have the same responses, participants view one incorrectly recognized stimulus, meaning no memory is associated with it, and another correctly recognized stimulus, making the word more likely to be from the detection state than a random word. The increased probability of one of the forced-choice test words coming from detection state makes it easier for participants to discriminate the two words. Therefore, percent correct of the forced-choice test should be higher than the single-item recognition.

The UVSD model may predict the opposite. Taking the O-O forced-choice test pair as an example. According to UVSD's assumptions, when a target and a lure are responded as "old",

both of them have memory strength falling on the right of the decision criterion. The lure has stronger memory strength than average lures, and thus more similar strength to targets than average lures. This can be seen from an illustration of how distribution means change when only parts on the right of the criterion are considered. In Figure 10, the distance between the two means is smaller in the second panel than the first one. As a result, it is more difficult to separate words from the two truncated distributions. Percent correct of the forced-choice test should be lower than single-item recognition.

The UVSD's prediction of lower percent correct for forced-choice test is in the same direction with the effects memory decay. Therefore, the above model fitting result favoring the UVSD model may be created by changes in memory across the two test types as opposed to demonstrating that this model is a better description of performance. However, this can be resolved in Experiment 2 because the two models predict opposite bias effects in O-O and N-N trials, and the bias predictions are not changed even if there is memory loss between the two trial types.

We also ran the simulation again with memory decay incorporated. A decreasing parameter *d* was randomly drawn from a uniform distribution between 0 and .2. After single-item recognition and before forced-choice test, the target distribution mean of the UVSD model decreased by *d*, and the criterion also decreased by half the decreasing parameter. For the 2HT model, the probability of detection result of both "old" and "new" responses decreased by *d*. Critically, we applied both models under the assumption that there was no memory change across test types, as we did for the empirical data. Thus, these simulations explore how results could be distorted if this assumption is incorrect.

Figure 11 and 12 show the simulation result. When the data were generated with the 2HT model, the fitting performance of the UVSD model did not change too much compared to when memory drop was not considered in the model. There were more large G^2 values for the 2HT fits compared to before. This made the median of the G^2 difference histogram now flipped over and was slightly smaller than 0, indicating a better fit by the UVSD model. When data were generated with the UVSD model, the fitting performance of the UVSD model was not affected as much as the 2HT model. The data were still better fitted by the UVSD model.

If a memory drop does occur during the period between single-item recognition and forced-choice test, Experiment 1 will not be able to discriminate the two models, since the UVSD model will win in both cases. Experiment 2 will address whether the advantage for the UVSD model in Experiment 1 is based only on a memory change between the two tests or indicates that the UVSD model better describes recognition memory.

2.5.2 Experiment 2

We obtained complete data from 45 participants of Experiment 2. We excluded 5 people's data from analysis. One person's single-item recognition percent correct was smaller than .6 (.53), one person had incomplete cycles, and another one did not make the initial guesses according to the bias information. The other two participants were dropped because they misunderstood the experiment procedure. The remaining 40 participants made guesses consistent with the payoff cues on .98 of trials: the proportion of "studied" guesses was .98 for the liberal trials and .02 for the conservative trials. Thus, participants closely followed the bias information on guess trials. The median RTs were 338 ms, 1,012 ms and 1,378 ms for the guess, single-item, and forced-choice trials, respectively.

The hit rate and false alarm rate of recognition test are shown in Table 3. The shift of bias observed in false alarm rates was .06 across two levels. Of all the 40 participants, 27 (68%) people had a bigger false alarm rate in the "studied safe" condition than the "not studied safe" conditions.

The 2HT and UVSD models make different predictions about the change of the forced-choice percent correct across bias levels. It can be seen in Figure 2, 3 and 4. From visual inspection, Figure 13 and 14 plotted from empirical data are very similar to Figure 3, which shows the UVSD model's prediction. The percent correct of O-O trials starts larger than that of the N-N trials, and from liberal to conservative, O-O trials' percent correct goes down. Also, the percent correct of the N-N trials goes up as the test becomes more conservative. This is exactly the opposite of the 2HT model's prediction.

The response percent correct of forced-choice trials was tested with a repeated measure 2 by 2 ANOVA in Experiment 2. There was a significant main effect of test type (O-O and N-N), $F(1, 39) = 64.96$, $p < .001$. The mean percent correct for the O-O and N-N test were .84 and .66, respectively. Interaction of bias and test type was also found to be significant, $F(1, 39) = 5.16$, $p = .03$. The bias effect was not significant, $F(1, 39) = 1.87$, $p = .18$. Mean percent correct for conservative and liberal trials were .77 and .74, respectively. The results are shown in Figure 13.

Considering the fact that not all participants in Experiment 2 showed bias in their single-item recognition test, the ANOVA was performed again on those who responded with biases (participants who had higher false alarm rates when "studied" was the safe response than when "not studied" was the sage response in single-item trials). There were 27 participants included in this analysis. The results are plotted in Figure 14. There was again a significant main effect of test type, $F(1, 26) = 40.23$, $p < .001$. Mean percent correct for the O-O and N-N test were .82

and .66, respectively. The bias effect did not reach significance, $F(1, 26) = 1.88$, $p = .18$, with means being .76 and .72 for conservative and liberal trials. The interaction between bias and test types was not significant this time, $F(1, 26) = 4.00$, $p = .06$. Despite the disappearance of the significance of interaction, Figure 13 and 14 resemble each other very well.

To further compare each model fit, we simulated performance data from the best fitting model parameters of each participant, and then plotted percent correct of the forced-choice trials the same as in Figure 13 and 14. Figure 15 is when performance data were generated with parameters of the 2HT model. Figure 16 is when the data were from the UVSD model. The simulation results followed the original models' predictions very well, and thus were very different from each other. For the 2HT model, the percent correct of O-O and N-N trials intersect around the no bias point. As bias became more conservative, the O-O trial's percent correct increased and N-N trial's percent correct decreased. When data were simulated with the UVSD model, the percent correct of O-O trials went down and N-N trials went up from liberal to conservative bias conditions. The percent correct difference between O-O and N-N trials became smaller from liberal to conservative, but did not reach zero. Still, data generated with the UVSD model followed the pattern of the actual data. In Figure 16, the simulated percent correct change from liberal to conservative of N-N trials is smaller than that of the actual data shown in Figure 13. Participants' percent correct in the liberal condition is lower than what was produced by the model. They had more difficulties in comparing memory evidence than the models.

Model fitting result is shown in Figure 17. Both UVSD and 2HT models were fitted to the combined single-trial and forced-choice data. The best fit parameters of both models are reported in Table 5. The summed G^2 followed a χ^2 distribution with 160 degrees of freedom, assuming the fitted model was the true description of the actual data. The summed G^2 value of

the UVSD model was 174, and was within the 0.95 quantile of the expected χ^2 distribution if UVSD is the true model. The summed G^2 value of the 2HT model was 688, and it exceeded the 0.95 quantile of the χ^2 distribution.

Among the 40 participants, 37 had smaller G^2 values for the UVSD fits than the 2HT (93%, $p < .001$ with binomial test). For individual G^2s, there were 28 (70%) out of 40 values that had significant χ^2 values ($p = 0.05$) with the 2HT model; no participants' value was significant with the UVSD model.

In Figure 17, the red curve in the first two panels outlined a χ^2 distribution with 4 degrees of freedom. A visual inspection reveals that the G^2 values of the UVSD model follow the χ^2 distribution better than the 2HT model does. Although there were two outliers whose G^2 values were significantly large, the median of the G^2 difference between the UVSD model and the 2HT model fits is still smaller than 0, suggesting a better fit of the UVSD model than the 2HT for the performance data.

CHAPTER 3

DISCUSSION

This study investigated the mechanism of recognition memory, with a focus on the debate between the continuous and discrete processes of recognition memory. We compared the most studied models from each field, i.e., the UVSD model from the continuous account and the 2HT model from the discrete account. Previous studies have tested these two models extensively, with techniques such as ROC functions and RT modeling, but consensus has not been reached. Our study further researched recognition memory by providing a new type of data: single-item recognition followed by the forced-choice testing on the same items.

Although a number previous studies have used the forced-choice testing method, or have conducted single-word recognition and forced-choice testing simultaneously to discriminate the two recognition memory models, our experiment combined the two tests together and made better use of the relationship between the types of responses. For instance, in previous studies, the single-item and forced-choice trials may occur intermixed with each other, but the two tasks' stimuli were separately sampled from the word list. Thus, the single-item and forced-choice tasks were essentially independent tests, and researchers were interested in the relationship between the two tests across participants. In contrast, for our experiments the item recognition and forced-choice tests were constructed from the same items. The forced-choice trials were created by pairing an incorrectly recognized target (or lure) with a correctly recognized lure (or target) from the single-item test. Each error in the single-item recognition task went through two decision making processes. The forced-choice responses were able to determine a stimulus' inner

cognitive state during the recognition process more precisely. To our knowledge, this is a novel way to make the forced-choice test trials.

With our way of making the forced-choice stimulus, the 2HT and UVSD models have different predictions on the pattern of forced-choice data with their different assumptions about error mechanisms. The 2HT model assumes that people's recognition judgments come from three discrete inner cognitive states: detect old, detect new and guessing. An item enters into the detect old or detect new state when some information that infallibly proves its previous presence or novelty is retrieved. For example, in a word recognition test, a participant may be tested on word that he remembers seeing on the word list. This word will be detected old and then given the "studied" response. He may also be tested on a verb but he knows every word on the word list was noun. This word is then detected new and given the "not studied" response. There will also be cases when no accountable information is retrieved about a test word, yet a response has to be made. This word then enters into the guessing state through where it is randomly given the "studied" or "not studied" response. An important feature of this model is that retrieval of infallible information will always lead to the correct response, so errors can only occur from the guessing state.

The UVSD model assumes continuous memory evidence of test stimuli, even for lures. Memory evidences are represented with two Gaussian distributions. Conventionally, the distribution of lures has a mean of 0 and standard deviation of 1. The distribution of targets has larger mean and standard deviation than the lures because of stronger memory evidence. The decision maker has a predetermined criterion set along the evidence axis. Any item with an evidence value smaller than the criterion will be called "not studied", and any items with larger evidence than the criterion will be called "studied". As part of the lure distribution has to extend

to the right of the criterion ("studied" area), and part of the target distribution extends to the left of the criterion ("not studies" area), when an item falling into such regions appear in recognition test, participants will be misled to make the incorrect decision. In other words, under UVSD's assumption, every test word is retrieve with some memory evidence. Errors occur not because of guessing failure, as the evidence has been successfully retrieved, but because the evidence is misleading. This is the point where the 2HT and UVSD model performances are able to be distinguished.

For the 2HT model, a recognition error is simply an unlucky guess. In the forced-choice test, when this word is paired with a correctly recognized item, the word will not be able to insert any useful information into the comparison. All the decision maker has to rely on is the other word in the forced-choice pair. If the other word is a detection result, the person will be able to pick the right word, because words from detection states are remembered with infallible information. If the other word of the pair is also a guessing result, then this person has no information to rely on. All he has to do is make another random guess. The fact that the pairing word is correctly recognized makes it more likely to have been detected than items overall, especially when participants are biased in the opposite direction, as there are fewer correct responses that are lucky guesses. In this case, the forced-choice test becomes easier than the single word recognition under assumptions of the 2HT model.

Unlike the 2HT model's prediction that only the correct response is affected when the response bias changes, under the UVSD model's prediction, both the correct and incorrect responses will change. The UVSD model assumes that an incorrect recognition response happens because the decision maker has indeed retrieved misleading information. When such word appears together with a correctly recognized one, their memory evidence will compete with each

other, making the participant more confused than in the single-item test. In this case, the forced-choice test is made to be more difficult than the single item recognition.

The 2HT model performed worse than the UVSD in our experiments because it over predicts the percent correct of the forced-choice test, due to its assumption of the error mechanism. More fundamentally, the reason why the two models make different predictions as well as different fitting results lies in the models' different assumptions about inner cognitive processes of decision making, namely, whether decisions come from comparisons between criteria and evidences drawn from a continuous strength value, or come from several discrete definite or guessing states. The continuous and discrete state account of memory information is the focus of the controversy.

In two experiments of this study, participants first studied a list of words and were tested on them. During the test, participants were presented with a word that either has been studied or not studied. They first went through the single-item test where they were asked to respond "old" to a word they thought was studied and "new" to a word they thought was not studied. In Experiment 1 participants responded to the single-item test without bias. In Experiment 2, they responded with two levels of bias. After every 10 or 12 such trials, participants were brought to a forced-choice test where they were presented with two words to which they had made the same responses: responses were both called "old" or both "new". Participants were notified that one of the decisions was incorrect and they had to choose a word that they think was studied before.

The forced-choice test provides an alternative way to test the continuous and discrete models. Simulations of Experiment 1 and 2 show that, when the forced-choice test data are generated from the UVSD model, UVSD will fit the data better than 2HT model, indicated by a smaller G^2. When the data are generated from the 2HT model, 2HT will fit the data better.

Modeling results of Experiment 1 and 2 have both supported the UVSD model. In Experiment 1, all 69 participants' performance was fitted by the two models. The UVSD model won over the 2HT for 55 cases. The G^2 distributions of the true model should follow a χ^2 distribution of 69 degrees of freedom. From visual inspection, the G^2 distribution of the UVSD model follows the χ^2 distribution better than the 2HT does. Also, the χ^2 test showed that the summed G^2 values of the 2HT model went past 0.95 quantile of the χ^2 distribution. In Experiment 2, the UVSD model fitted 37 participants better out of 40. It also performed better than 2HT model in the χ^2 test.

Result of the ANOVA test in Experiments 2 was also informative. According to the 2HT model's prediction of the forced-choice data, when the guessing parameter decreases (decision becomes more conservative), one would expect the percent correct of the old-old test to increase and the percent correct of the new-new test to decrease. There will be an interaction between the guessing bias and test types, like what is shown in Figure 2. However, this is not observed in the data. On the contrary, the effect in Figure 13 was very much like what was shown in Figure 3, where the simulation data were generated with the UVSD model. In Figures 13 and 3, the percent correct of old-old test was higher than that of the new-new test. As the decision criterion increases, percent correct of old-old test slightly goes down and the new-new test goes up.

One benefit of the forced-choice paradigm is that the 2HT and UVSD models are distinguishable without a bias manipulation. It also does not require specific features from the 2HT or UVSD models, so it can be used to test other models of recognition memory, such as the single high-threshold model (1HT) and low threshold (LT) models of the discrete account, and the dual-process signal detection (DPSD) model of the hybrid account. The model's prediction can be determined by either mathematical analysis or simulation. As long as the tested models make different predictions about the forced-choice data, their performances can be distinguished.

Previous studies have tried to distinguish the discrete state and signal detection theory models of recognition memory with techniques such as the ROC functions and RT modeling. There was also research based on the forced-choice test. Jang et al., (2009) asked participants to respond to single-item and 2AFC trials with 6-level confidence rating. There was no connection between the two test trials. During the model fitting, data from each test were fitted with parameters predetermined according to the relationship of the two tests. Parks & Yonelinas (2009) designed a 4AFC task with two responses. The first response was to choose one word out of four as the target word. The second response contained two ordered responses. The first one was to choose a word that most likely to be the target, and the second response chose the next most likely target word. Memory models were also fitted to the pattern of the two series responses separately.

In previous studies, the forced-choice test stimuli were directly sampled from the word list. During tests, the forced-choice test was usually accompanied by some response manipulations, such as confidence rating or ranking. In our experiments, the 2AFC test pairs were made based on the result of the single-item recognition tests. Without further manipulation, the 2HT and UVSD models were distinguishable with our tests (Experiment 1). Experiment 2 was conducted to address the concern of memory decay. Two levels of biases were introduced to the single-item and forced-choice tests. The 2HT and UVSD models made different predictions about the forced-choice percent correct on New-New and Old-Old trials. Model fitting results favored the UVSD model in Experiment 1, so did the ANOVA and modeling results in Experiment 2.

The present study has a few limitations. First, there was big variability in the number of forced-choice trials completed across participants. Conventional tests such as ANOVA are not

sensitive to this variance. Further analysis can use hierarchical Bayesian models to account for the individual differences. Further study can also make the forced-choice trials appear immediately after the single-item trial to avoid memory decay from happening. To prevent participants from guessing the correct response in this design, every single-item trial will be followed by a forced-choice trial.

Our study aims to distinguish two major models of recognition memory: the 2HT model and the UVSD model. The experiments combined the single-item recognition and the forced-choice test. To our knowledge, this is a novel way to test recognition memory. Two experiments provided strong support for the UVSD model.

Table 1: An example of a cycle of single-item and forced-choice trials

Single-item Trail			Subsequent Forced-Choice Trial	
Word	Attribute	Response	Trial type	Word pair
Compassion	Target	Old		
Restaurant	Lure	Old	O-O	Restaurant-Way
Lady	Lure	New		
Flight	Lure	New		
Obedient	Lure	New		
Spectacular	Target	New	N-N	Lady-Spectacular
Way	Target	Old		
Alien	Target	Old		
Potency	Target	Old		
Disaster	Target	New	N-N	Disaster-Flight

Table 2: Descriptive statistics of Experiment 1

Single-item recognition			Forced-choice test
HR	FAR	Percent correct	Percent correct
.64 (.02)	.25 (.01)	.69 (.007)	.64 (.01)

Table 3: Best fit parameter values of Experiment 1

Measure	2HT			UVSD		
	Do	Dn	g	μ	σ	λ
Mean	.34 (.02)	.24 (.02)	.39 (.02)	1.06 (.06)	1.20 (.04)	.664 (.05)
Upper CI	.38	.28	.43	1.18	1.28	.76
Lower CI	.29	.21	.35	.94	1.13	.57

Table 4: Hit (HR) and false alarm rates (FAR) in Experiment 2. Standard errors are in parenthesis.

Performance measure	Single-item recognition		Forced-choice test	
	Liberal	Conservative	Liberal	Conservative
HR	.75 (.02)	.82 (.02)	.63 (.03)	.70 (.02)
FAR	.20 (.02)	.14 (.01)	.15 (.02)	.16 (.02)

Table 5: Best fit parameter values of Experiment 2. l indicates liberal and c indicates conservative

Measure	2HT				UVSD			
	Do	Dn	g_l	g_c	μ	σ	λc	λ_l
Mean	.66 (.03)	.38 (.03)	.37 (.03)	.24 (.02)	2.64 (.46)	1.67 (.30)	1.22 (.07)	.94 (.08)
Upper CI	.72	.45	.42	.29	3.57	2.26	1.37	1.11
Lower CI	.60	.31	.31	.20	1.72	1.07	1.07	.78

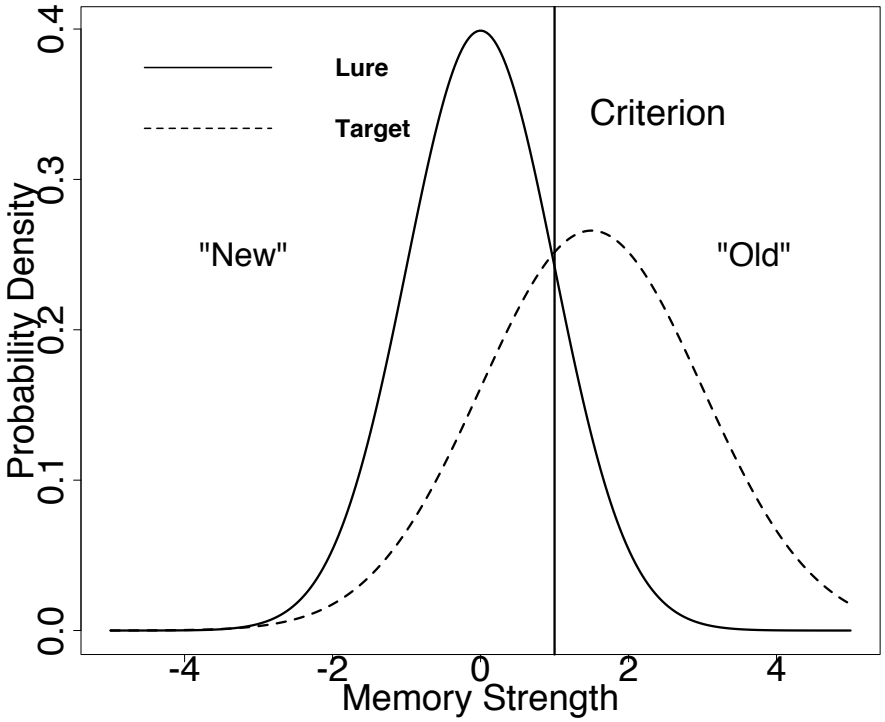

Figure 1: SDT model for recognition memory

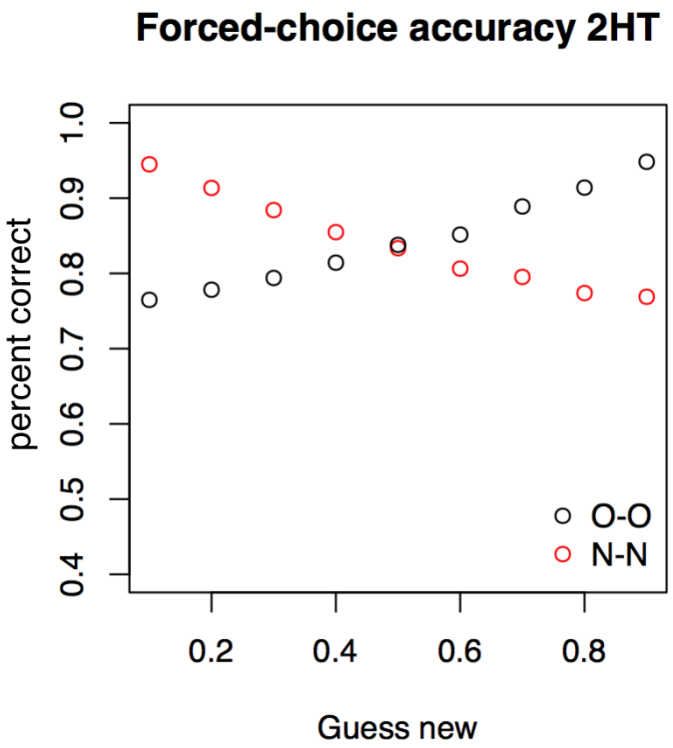

Figure 2: Illustrations of the forced-choice test percent correct as the guessing parameter changes. Data were simulated with 1,000 trials. Detect old and detect new parameters were .5.

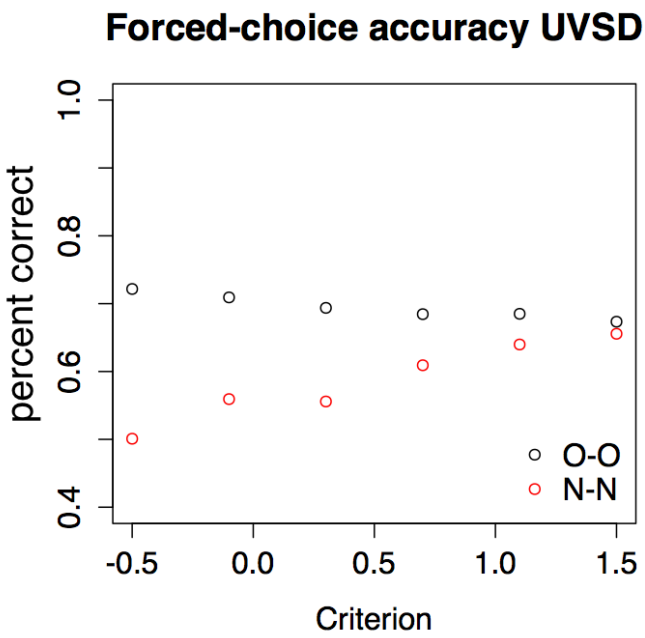

Figure 3: Illustrations of the forced-choice test percent correct as the decision criterion changes. Data were simulated with 1,000 trials. The mean and standard deviation of the target distribution were 1 and 1.2.

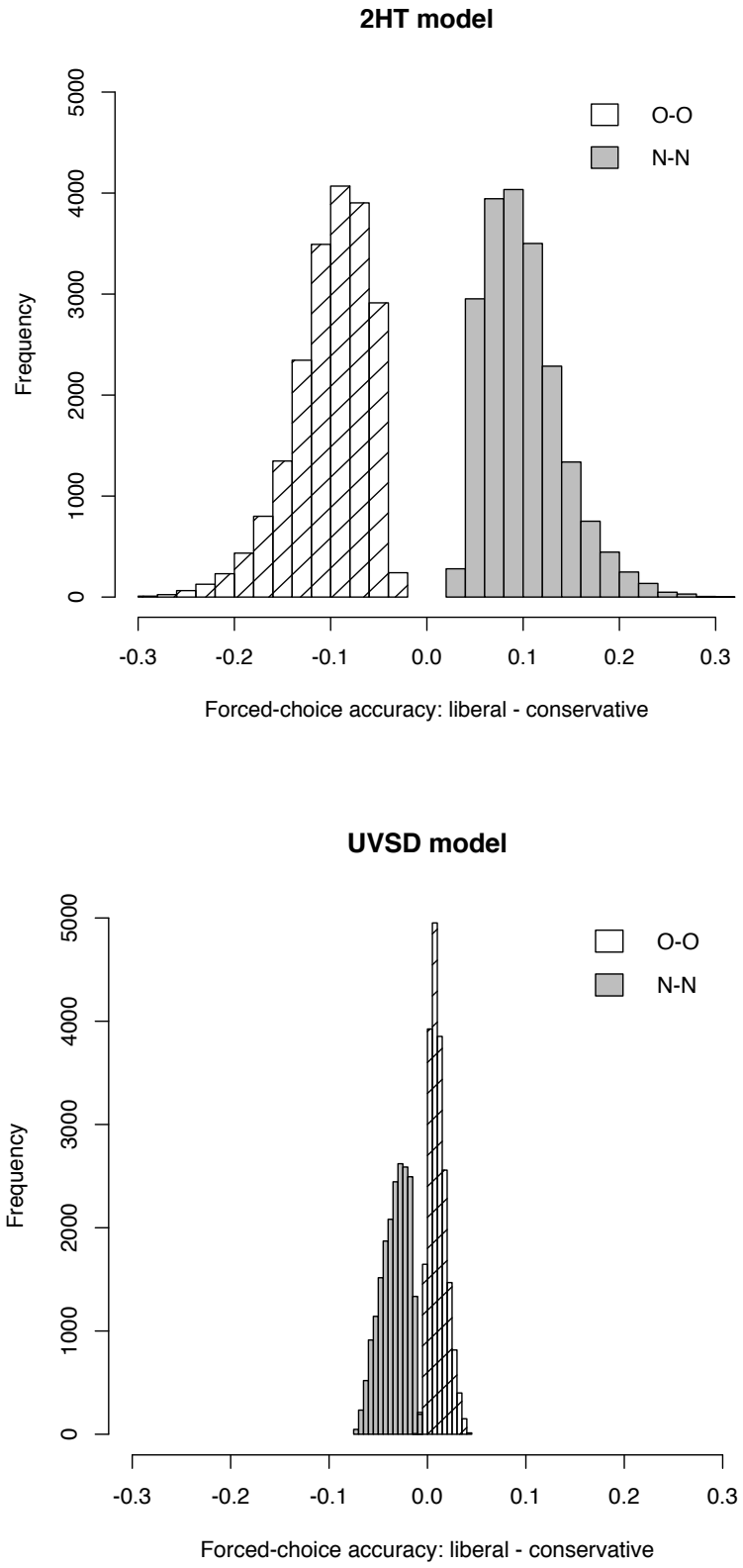

Figure 4: Predicted bias effects with randomly sampled parameter values.

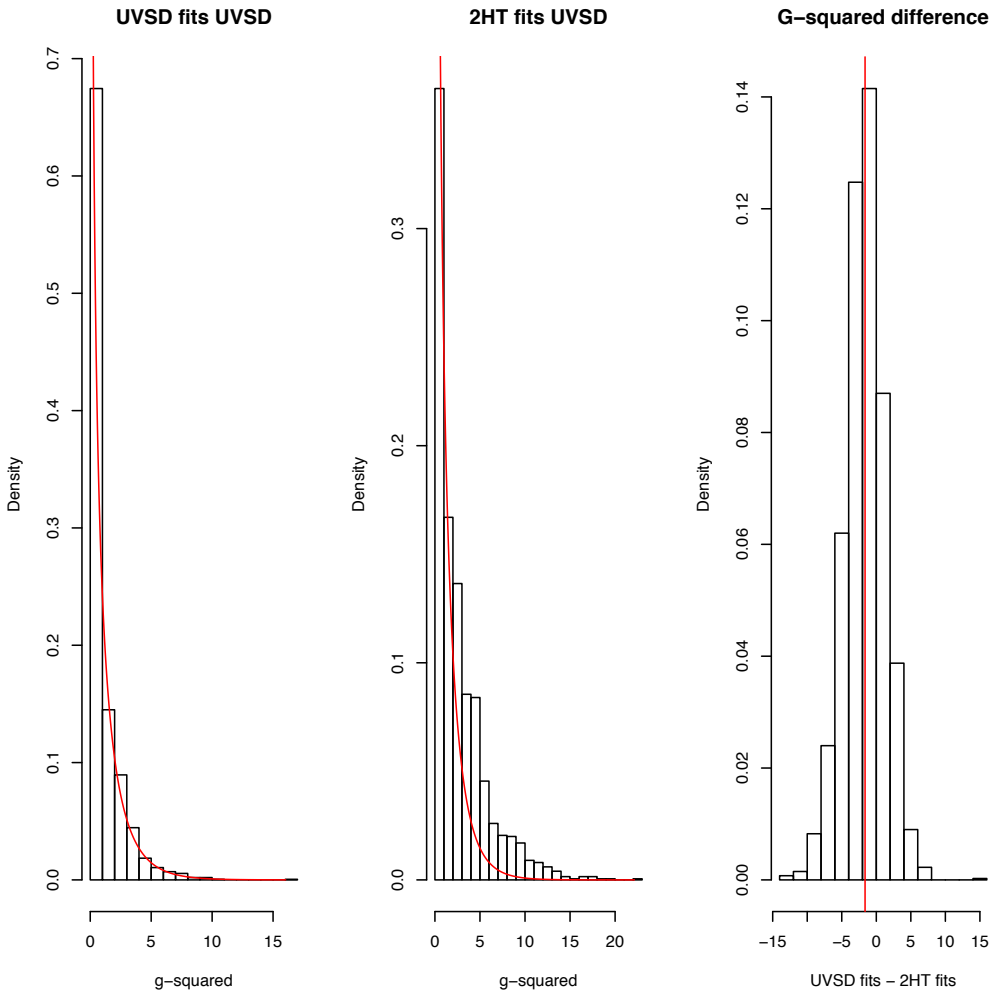

Figure 5: Simulation results of Experiment 1 when data were generated with the UVSD model. The first two panels show the G^2 histogram of the UVSD and 2HT fits respectively. The third panel plots the difference of G^2 between the two models fits. The red lines in the first two panels were the χ^2 distributions that the histograms should follow. The red vertical line in the third panel indicates the median of the histogram.

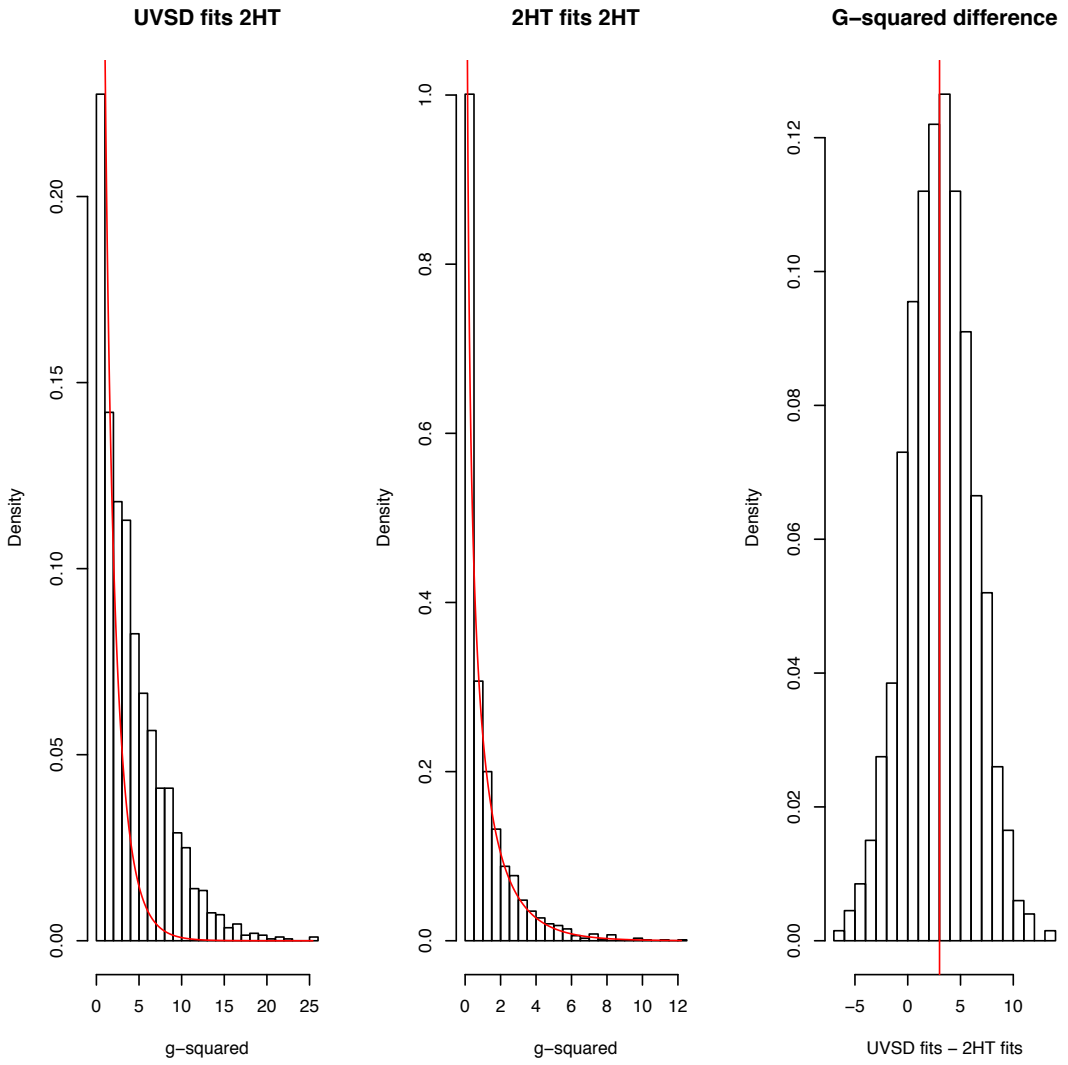

Figure 6: Simulation results of Experiment 1 when data were generated with the 2HT model. The first two panels show the G^2 histogram of the UVSD and 2HT fits respectively. The third panel plots the difference of G^2 between the two models. The red lines in the first two panels were the χ^2 distributions that the histograms should follow. The red vertical line in the third panel indicates the median of the histogram.

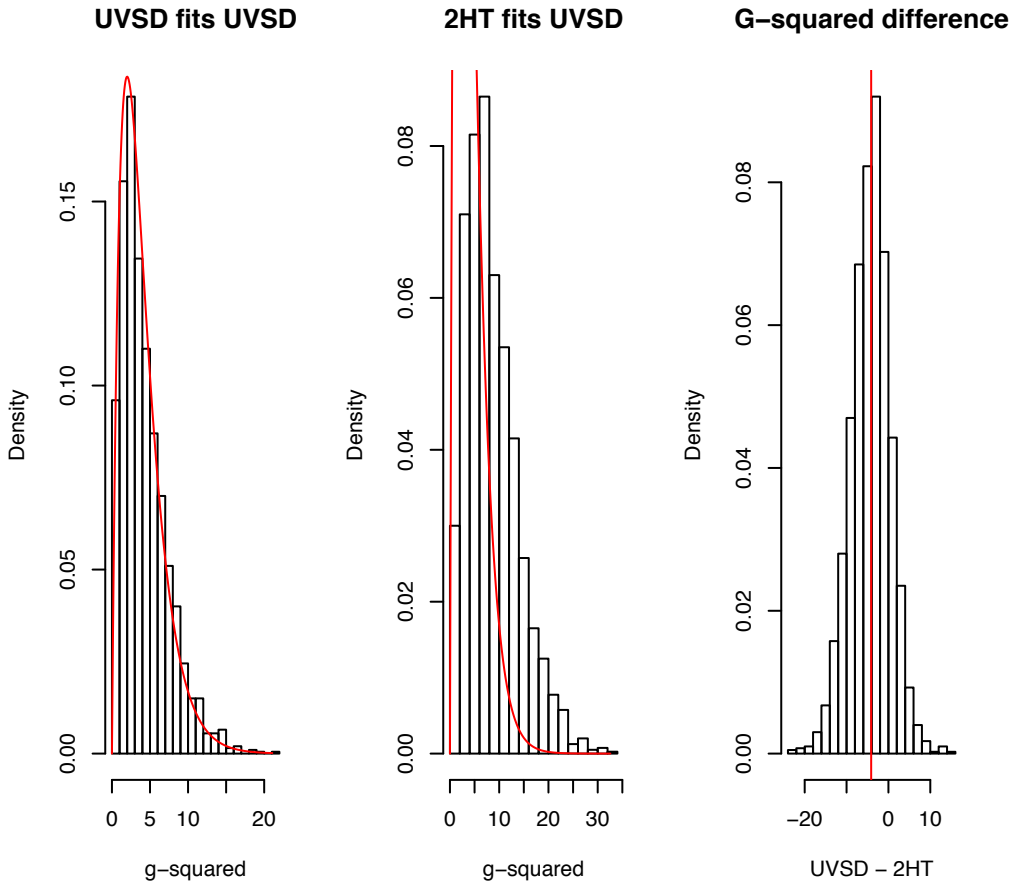

Figure 7: Simulation results of Experiment 2 when data were generated with the UVSD model. The first two panels show the G^2 histogram of the UVSD and 2HT fits respectively. The third panel plots the difference of G^2 between the two models. The red lines in the first two panels were the χ^2 distributions that the histograms should follow. The red vertical line in the third panel indicates the median of the histogram.

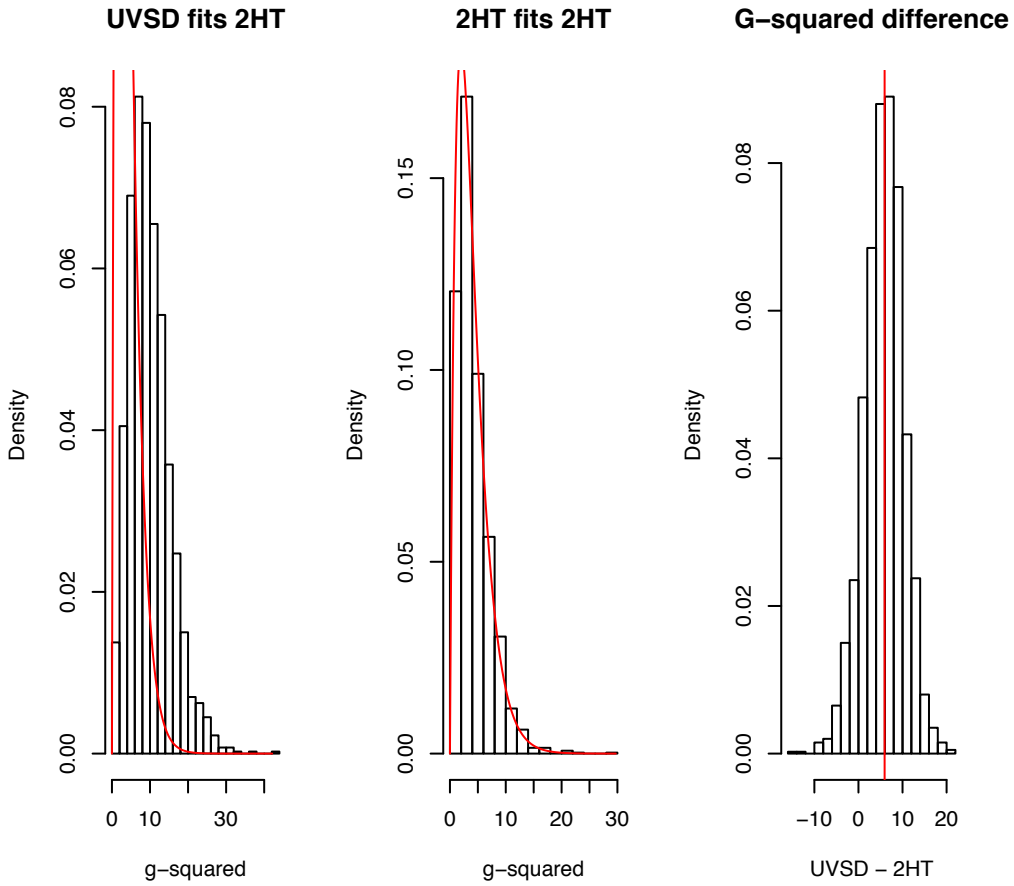

Figure 8: Simulation results of Experiment 2 when data were generated with the 2HT model. The first two panels show the G^2 histogram of the UVSD and 2HT fits respectively. The third panel plots the difference of G^2 between the two models. The red lines in the first two panels were the χ^2 distributions that the histograms should follow. The red vertical line in the third panel indicates the median of the histogram.

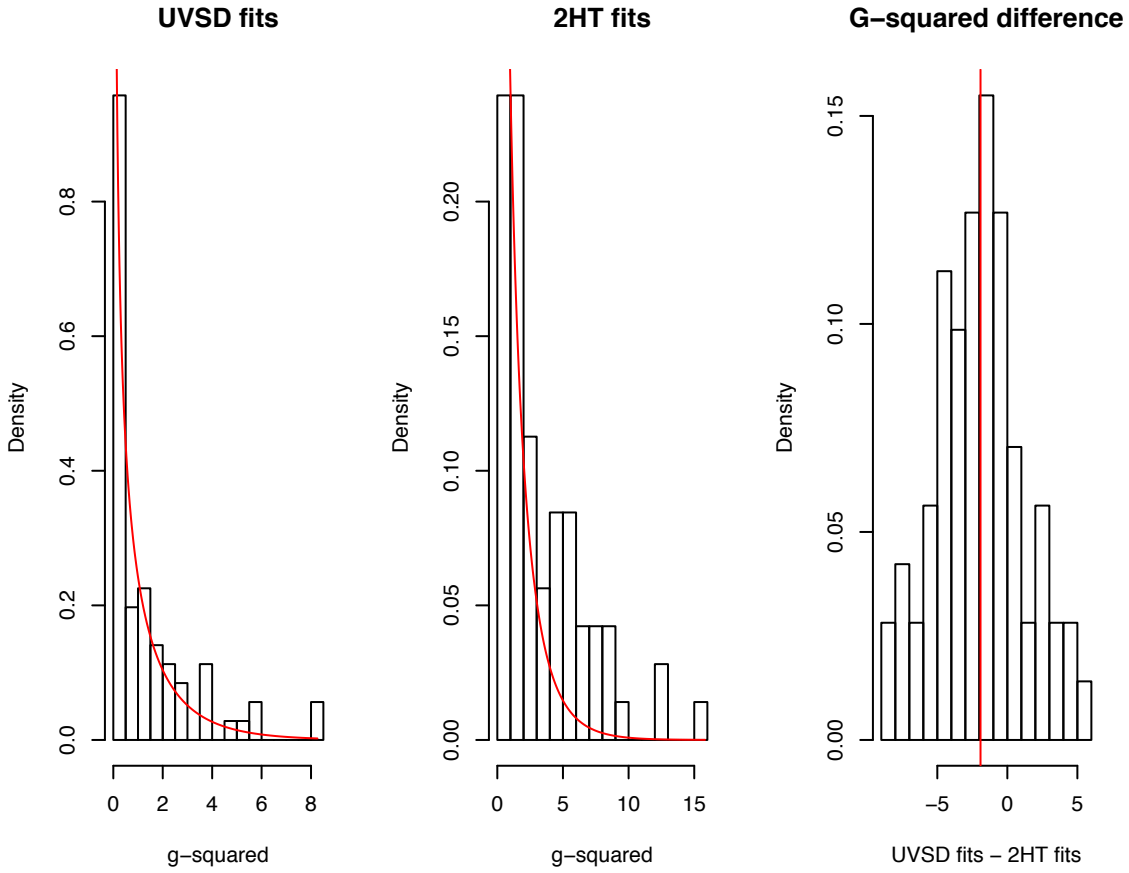

Figure 9: Model fitting results of Experiment 1. The first two panels show the histogram of G^2 for each model fits. The third panel shows the difference of G^2 between the two fits. The red lines in the first two panels were the χ^2 distributions that the G^2 distributions should follow. The red vertical line in the third panel indicates the median of the histogram.

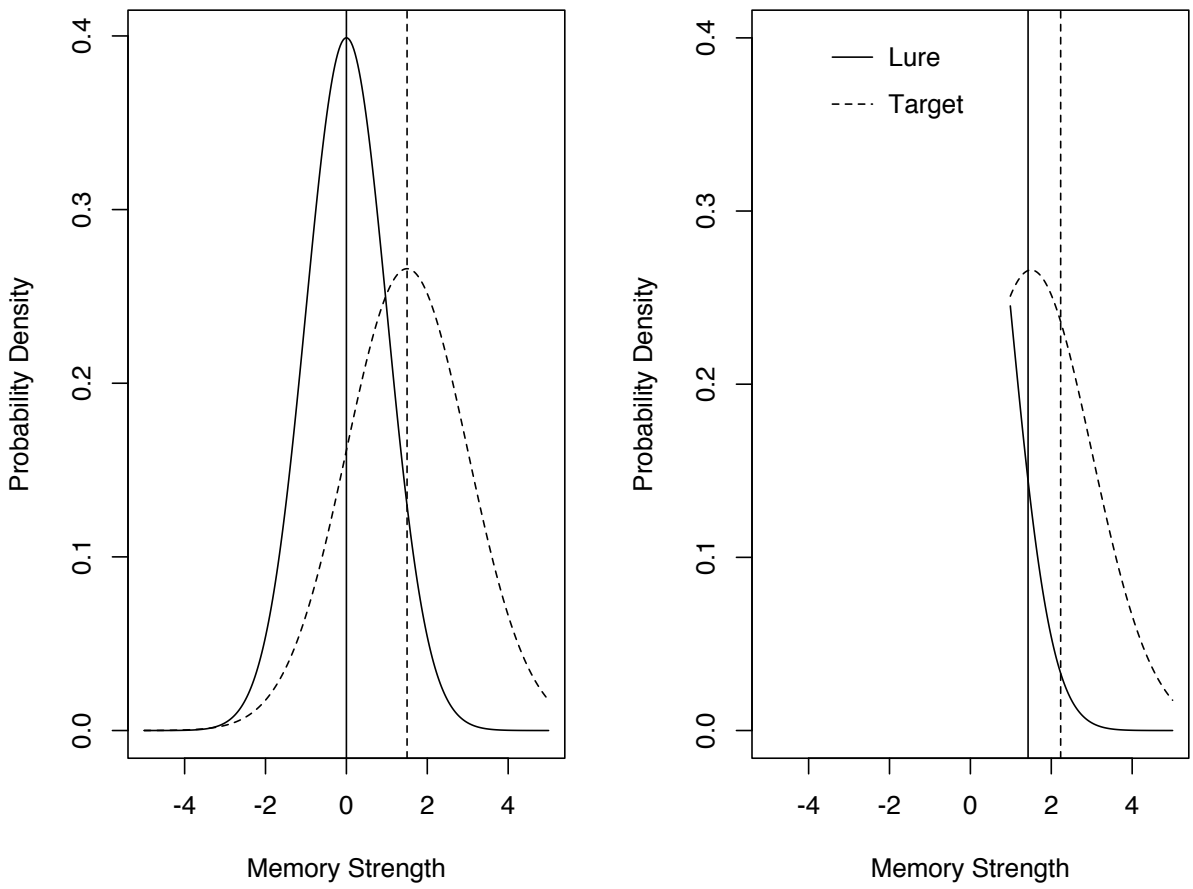

Figure 10: The distributions of target and lure. The two vertical lines plot the means. The left panel shows the entire distributions and their means. The right panel shows only the "old" stimuli and their means.

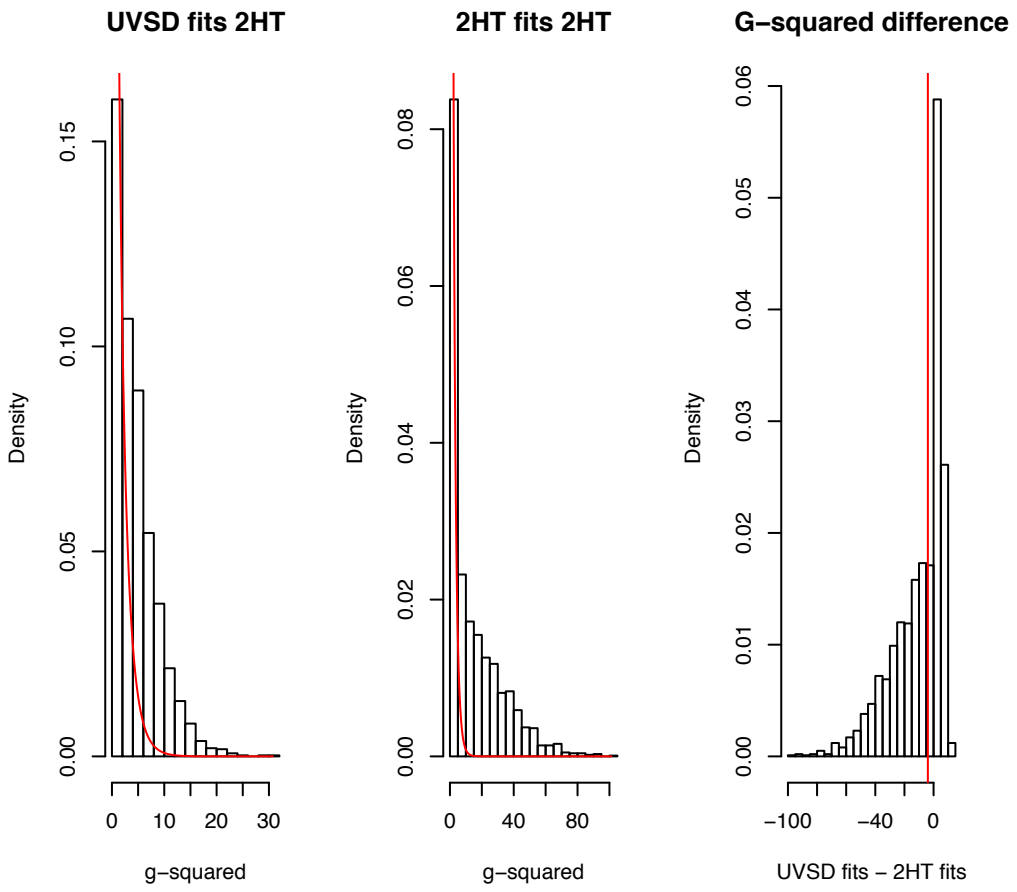

Figure 11: Simulation result when memory drop was considered. Data were generated with the 2HT model. The red lines in the first two panels were the χ^2 distributions that the G^2 distributions should follow. The red vertical line in the third panel indicates the median of the histogram.

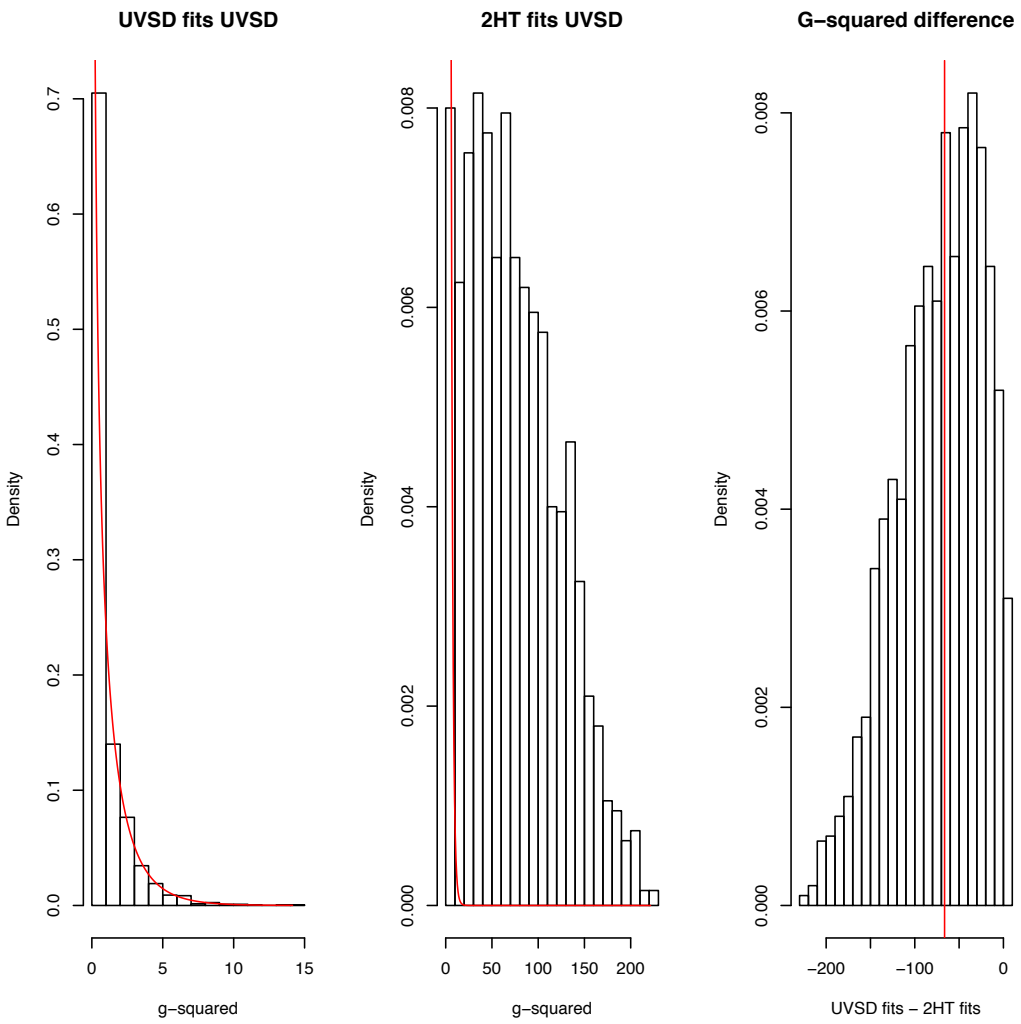

Figure 12: Simulation result when memory drop was considered. Data were generated with the UVSD model. The red lines in the first two panels were the χ^2 distributions that the G^2 distributions should follow. The red vertical line in the third panel indicates the median of the histogram.

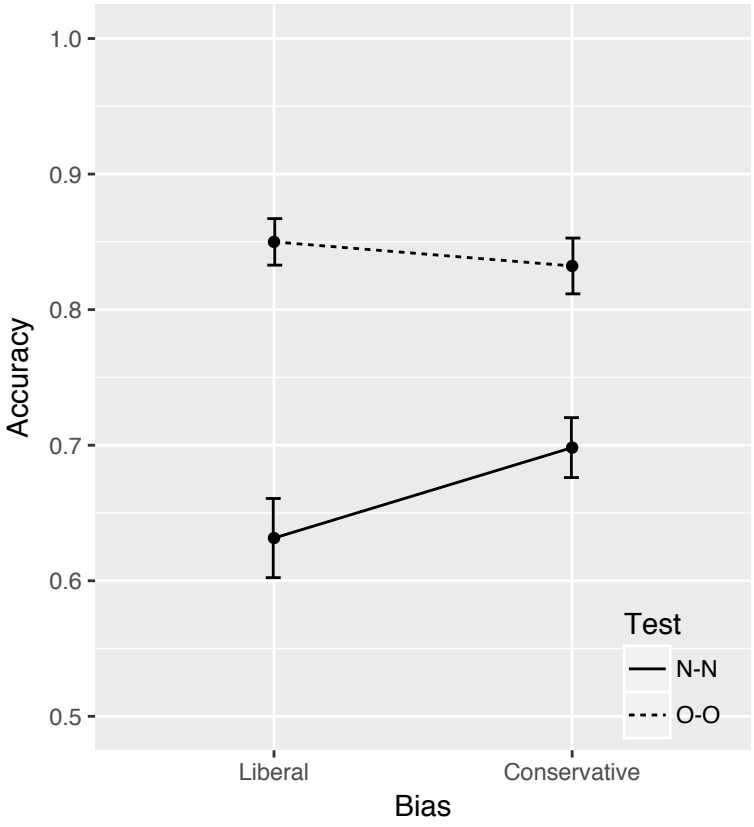

Figure 13: Percent correct of the forced-choice test of Experiment 2.

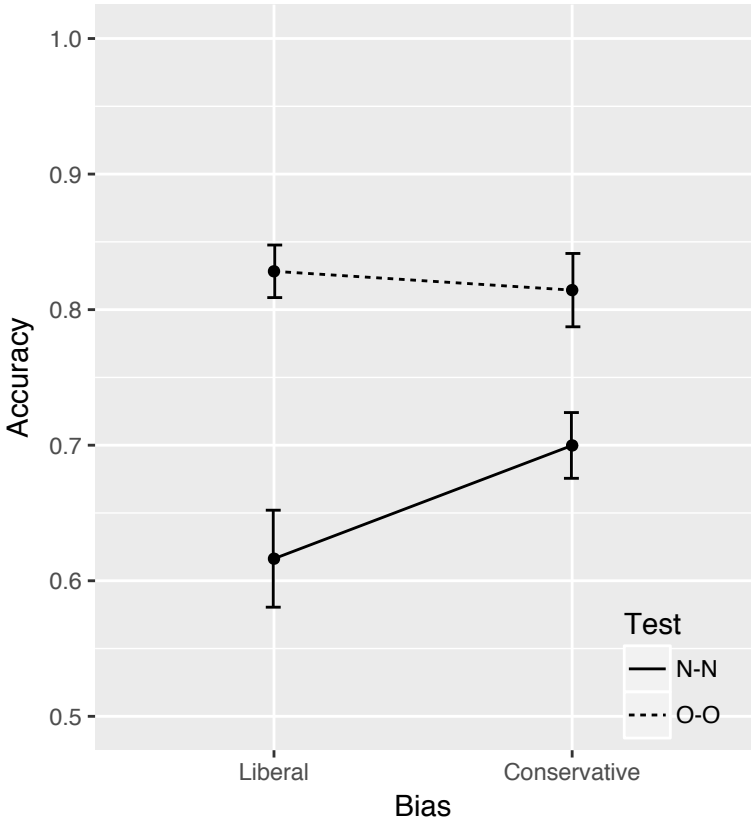

Figure 14: Forced-choice test percent correct of participants showed bias effect in single-item recognition trials.

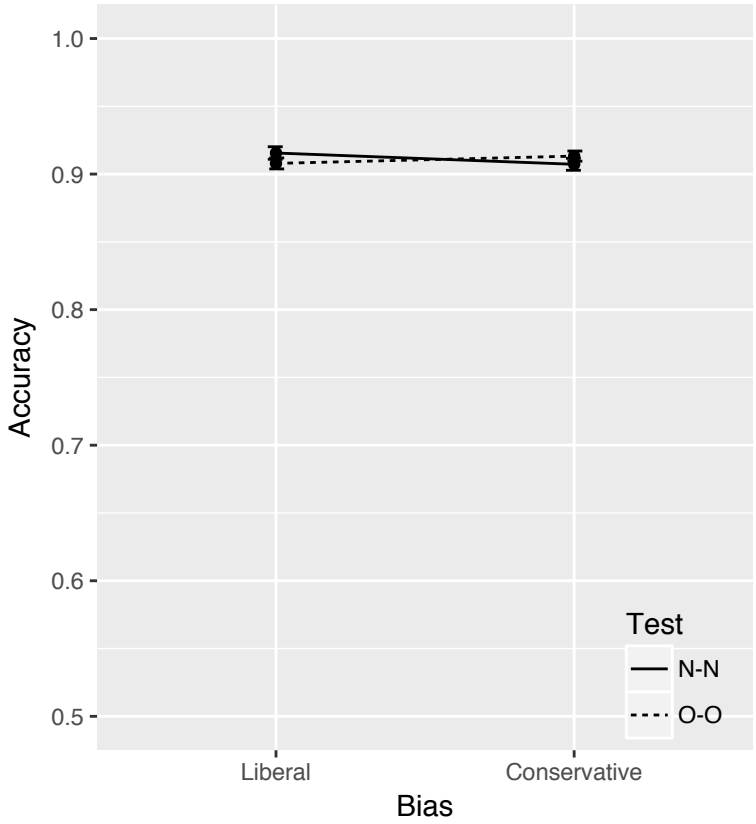

Figure 15: Percent correct of the forced-choice trials. Data were simulated from the best fitting 2HT model parameters of each participant.

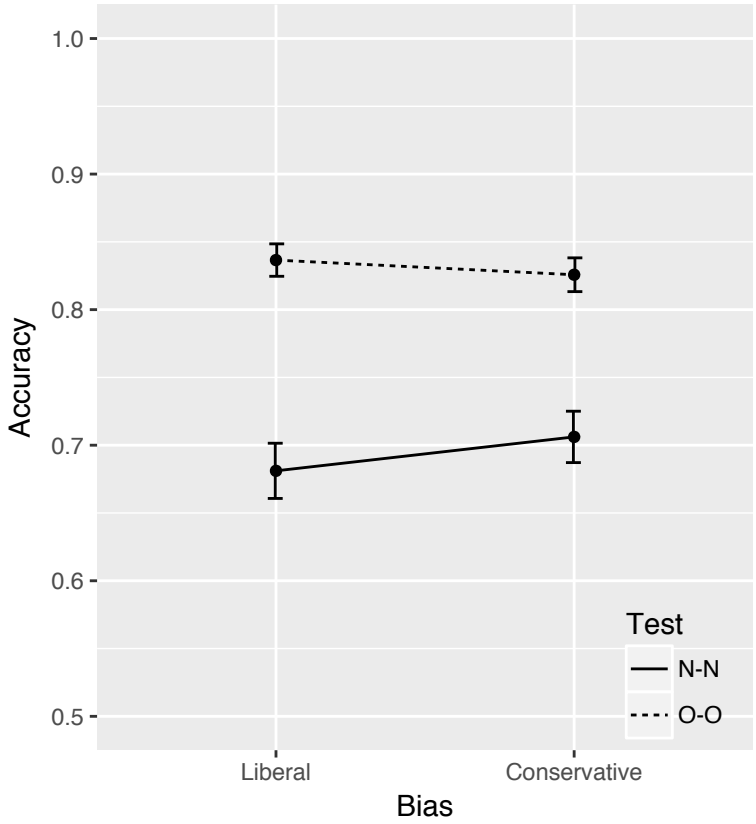

Figure 16: Percent correct of the forced-choice trials. Data were simulated from the best fitting UVSD model parameters of each participant.

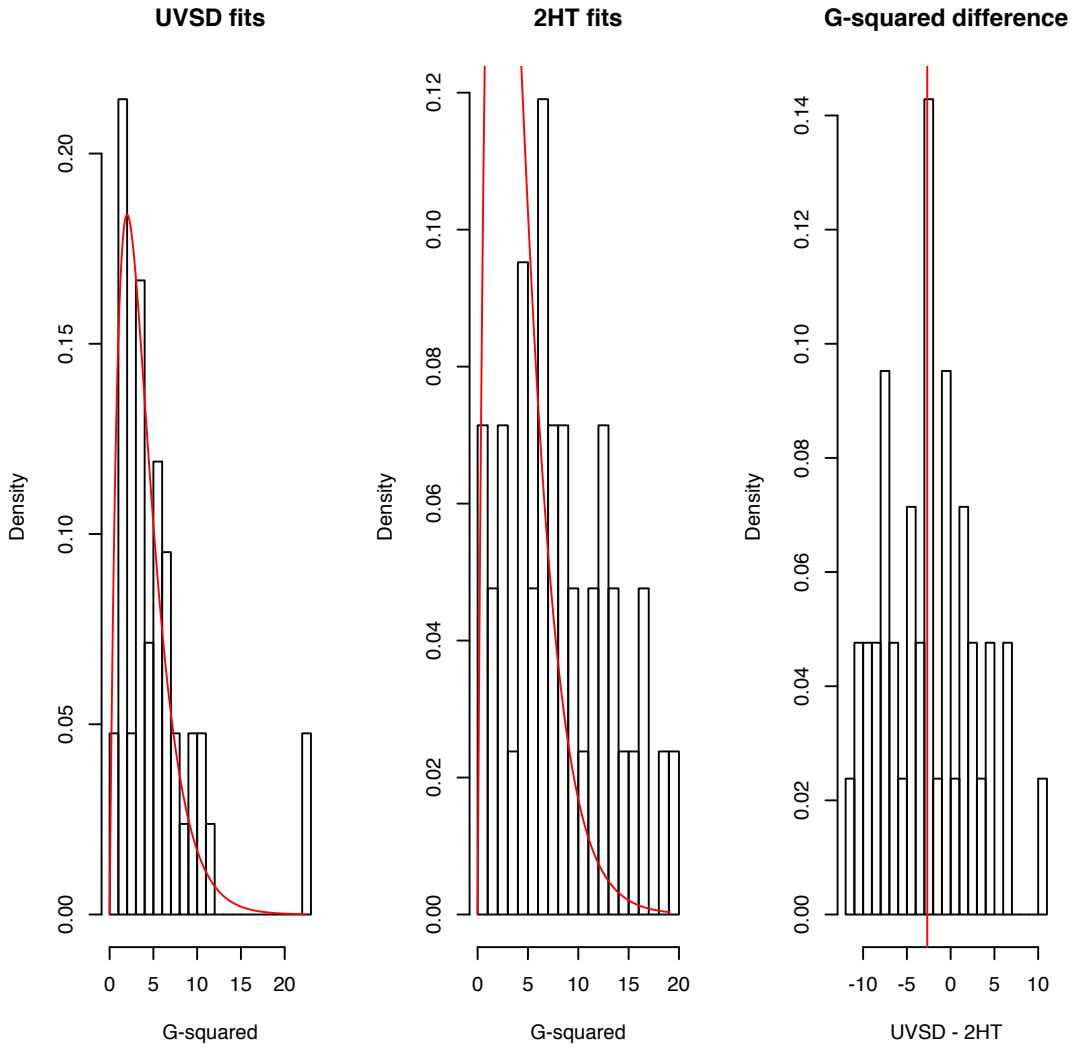

Figure 17: Model fitting results of Experiment 2. The first two panels show the histogram of G^2 for each model fits. The third panel shows the difference of G^2 between the two fits. The red lines in the first two panels show the χ^2 distributions that the G^2 distributions should follow. The red vertical line in the third panel indicates the median of the histogram.

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Akademiai Kiado, Budapest.

Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, *66,* 912-918.

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

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977-990.

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

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psycho- physics*. Oxford, England: Wiley.

Hu, X. (2001). Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology, 45*, 603-634.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138*, 291-306.

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first-and second-choice responses. *Journal of Mathematical Psychology, 55*, 251-266.

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1795-1804.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review, 20, 693*-719.

Kellen, D., Singmann, H., Vogt, J., & Klauer, K. C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology, 62,* 40-53.

Kroll, N. E., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General, 131*, 241-254.

Kucera, H. & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70,* 61–79.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 380–387.

Onyper, S. V., Zhang, Y. X., & Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General, 139*, 341-364.

Parks, C. M., & Yonelinas, A. P. (2009). Evidence for a memory threshold in second-choice recognition memory responses. *Proceedings of the National Academy of Sciences, 106*, 11515-11519.

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139,* 1173-1203.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 109,* 14357-14362.

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

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological review, 99*, 518-535.

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

Smith, D. G., & Duncan, M. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 615-625.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34–50.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64*, 1-34.

Swets, J. A. (1961). Is there a sensory threshold? *Science, 134,* 168–177.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192-196.

Wixted, J. T. (2007). Dual-Process Theory and Signal-Detection Theory of Recognition Memory. *Psychological Review*, 114, 152-176.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, 133, 800-832.