Bayesian Analysis of Recognition Memory: The Case of the List-Length Effect

Simon Dennis
Department of Psychology
Ohio State University

Michael D. Lee
Department of Cognitive Sciences
University of California, Irvine

Angela Kinnell
School of Psychology
University of Adelaide

Recognition memory experiments are an important source of empirical constraints for theories and models of memory. Unfortunately, standard methods for analyzing recognition memory data have problems that are often severe enough to prevent clear answers being obtained. A key example is whether longer lists lead to poorer recognition performance. The presence or absence of such a list length effect is a critical test of competing item- and context-noise based theories of interference, but remains an unresolved empirical issue, largely because of the weaknesses of the standard analysis. In this paper, we develop a new Bayesian method of analysis that overcomes the problems. We report data from a new recognition memory experiment that manipulates list length, as well as the better understood manipulation of word frequency, and present both standard and Bayesian analyses of the data. The comparison of the two methods allows us to highlight the advantages of the Bayesian approach in inferring the values of psychologically meaningful variables, and in choosing between models representing different theoretical assumptions about memory.

In a typical yes/no recognition memory task, subjects are asked to study a list of items and then decide whether or not each of a set of test items appeared on the study list. This task has been a touchstone for understanding episodic memory (Glanzer & Adams, 1985; Ratcliff, Clark, & Shiffrin, 1990), and has provided important constraint for a series of memory models (Gillund & Shiffrin, 1984; Murdock, 1982; Eich, 1982; Hintzman, 1986; Humphreys, Bain, & Pike, 1989; Shiffrin & Steyvers, 1997; McClelland & Chappell, 1998; Clark & Gronlund, 1996; Dennis & Humphreys, 2001). Recently, however, there has been debate concerning the primary source of interference in recognition memory paradigms. Logically, interference can arise either from the other items that appear in the study list, or from the other contexts in which a test item has appeared, or from both sources (Humphreys, Wiles, & Dennis, 1994).

A critical empirical test of these competing theoretical positions involves the presence or absence of list length effects. If item noise is the primary source of interference, recognition should be poorer for longer study lists than for shorter ones. If context is the primary source of interference,

Address correspondence to: Simon Dennis, Department of Psychology, Ohio State University, Columbus, OH 43201, U.S.A. Telephone: +1 614 292 2229. Electronic Mail: simon.dennis@gmail.com

changes in the length of the study list should not change recognition performance. Currently, there is no consensus on whether or not a list length effect is observed empirically, in part, because there are a number of confounds that could produce artifactual list length effects.

The most obvious of these is the retention interval. If one presents a study list followed immediately by test, then retention interval will be longer for the long list. There are two ways in which retention interval can be equated. In a retroactive condition, filler activity is added after the short list and only items from the start of the long list are tested (see Figure 1). In a proactive condition, filler activity is added before the short list and only items from the end of the long list are tested.

Using the retroactive design, Schulman (1974) found no list length in a forced choice test. Bowles and Glanzer (1983) did not analyze the retroactive condition separately from the proactive condition, but the difference in the proportion correct between short and long conditions was small (0.033) Also, in the third experiment of Murnane and Shiffrin (1991) the effect of length was not significant. In contrast to previous work, Gronlund and Elam (1994) did find a significant effect of length using a retroactive design. In this experiment, intentional instructions were employed and we will argue below that rehearsal could have been a factor.

In experiments employing proactive designs, the effects of length have been more robust. Bowles and Glanzer (1983) found a difference of 0.068 in the proactive condition, and overall found a significant effect of length. Underwood (1978) used a forced choice test and found an effect of length, as did Ohrt and Gronlund (1999). Underwood, citing the stability of word difficulty across list lengths and the lack of cumulative proactive interference in other recognition paradigms, argued against the direct involvement of proactive interference in recognition.

Rather, Underwood (1978) suggested that list length effects in proactive designs were caused by a lack of attention. In long lists, subjects must maintain attention throughout the list. The items tested are those at the end of the list, which are the ones most likely to be affected by attentional lapses. In contrast, in short lists all items effectively appear at the start of the list. In the Bowles and Glanzer (1983) study, the long list contained 240 words. In the Underwood (1978) study, the long list contained 80 words and in the Ohrt and Gronlund (1999) study, the long list contained 82 words. In all three cases, words were presented for 1.5 to 2.0 seconds under intentional learning instructions, but with no specific processing requirements and no way of ensuring that attention was maintained. Lapses of attention seem likely under these conditions, particularly in the case of Ohrt and Gronlund (1999) in which subjects participated in four 50-minute sessions.

A third potential confound is rehearsal. In the retroactive condition, a filler task is introduced between study and test. If subjects devote any of this time to rehearsing the studied items then performance in the short list will be superior to that in the long list both because there is more time to rehearse the short list and because any rehearsal that subjects might engage in under the long list conditions will be spread across more items and quite probably be focused on later items that will not be tested. Both experiments conducted by Gronlund and Elam (1994) involved intentional conditions, which increases the likelihood of rehearsal.

The fourth potential locus of an artifactual list length effect, and the one on which we will focus in this paper, is contextual reinstatement. Episodic recognition necessarily involves the use of both an item and a context cue (Humphreys et al., 1994). In the retroactive design, subjects are either tested immediately in the long condition, or after the filler task in the short condition. After the long list, as far as the subject is aware, the current context can be used to initiate retrieval. However, in the case of the short list the current context focuses on the filler task, and so the subject is likely to reinstate the context of the study list so as to isolate the relevant study episode. To the extend that context drifts during the presentation of the long list, the end of list context may not be an efficacious cue for items that were presented at the start of the list and hence performance in the long list will suffer.

Controlling for the factors outlined above Dennis and Humphreys (2001) argued that, for verbal stimuli, context is the primary source of inter-

ference, and presented empirical evidence consistent with the absence of a list length effect. Cary and Reder (2003) contested this conclusion, and presented empirical evidence consistent with a list length effect.

There were a number of differences between the two studies that could explain the different results. Cary and Reder (2003) only analyzed the combined proactive and retroactive results. As we argue above, list length effects in proactive designs have typically been larger than in retroactive designs, perhaps because of the effects of attention as suggested by Underwood (1978). Secondly, Cary and Reder (2003) employed the remember know procedure which requires subjects to attempt to recall specific aspects of the study episode. In recall, the existence of a list length effect is not disputed, so it is possible that recall is contaminating the results in a way that did not occur in the Dennis and Humphreys (2001) experiments, which used yes/no recognition. Thirdly, Cary and Reder (2003) employed a much shorter period (two minutes) between the end of the long list and test than did Dennis and Humphreys (2001, eight minutes). It is possible that this shorter period was not sufficient to compel subjects to engage in contextual reinstatement in the long condition.

The source of interference is a fundamental aspect of understanding memory phenomena, and so this debate is crucial to the development of models of recognition memory. Unfortunately, the appropriate way to analyze recognition data has been a controversial topic (Banks, 1970; Lockhart & Murdock, 1970; Snodgrass & Corwin, 1988), because the methodology that is used standardly has a number of undesirable properties. These fall into two main classes: Those related to the application of Signal Detection Theory (SDT), and those related to the application of standard methods for statistical inference. In this paper, we accept the standard SDT assumptions, but develop a Bayesian framework for understanding recognition memory performance that improves how the model can be related to experimental data. In particular, we tackle both issues of parameter estimation caused by the standard use of frequentist methods, and issues of model selection and evaluation caused by the standard use of Null Hypothesis Significance Testing (NHST).

We start by describing a new recognition memory experiment. We outline the method of analysis that would commonly be used in the recognition literature and, by applying it to the new data, describe its deficiencies. We then introduce and apply the Bayesian approach to the same data, and contrast its findings to the standard results. Finally, we relate these findings to our theoretical understanding of recognition memory.

Experiment

In this experiment, we investigated the impact of contextual reinstatement on the list length effect by manipulating whether additional filler activity was introduced after the long and short lists.

Method

Subjects. Forty eight Psychology I students from the University of Adelaide participated in exchange for course credit. The 10 male and 38 female subjects ranged in age from 16 to 42 years (mean = 19.71, standard deviation (SD) = 4.73). The sample size was equivalent to that in Dennis and Humphreys (2001) study and larger than in Cary and Reders (2003) study.

Design. A 2 x 2 x 2 factorial design was used. The factors were list length (short or long), condition (filler or no filler activity) and word frequency (high or low). All comparisons were within subjects.

Materials. Two hundred and eighty words were selected from the Sydney Morning Herald Word Database for use in this experiment. Half of the words were five letters and the rest were six letters long. Half of the words were of high frequency, defined as occurring 100 – 200 times per million words, and the remainder were low frequency words, of between one and four occurrences per million. The criterion for each frequency definition was consistent with the study of Dennis and Humphreys (2001). Each study and test list comprised an equal number of randomly ordered five and six letter, and high and low frequency words. Words were randomly assigned to conditions and

no word appeared twice in the study, with the obvious exception of targets.

Procedure. Upon arrival, subjects were given an overview of the experiment but were not told that there would be lists of different lengths or words of different frequency. Prior to the first study list, the instructions for a sliding tile puzzle activity were displayed on the screen and subjects completed a 30 second trial of the puzzle. Four lists were presented to subjects, two short and two long. As in Cary and Reder's (2003) study, short study lists were 20 words long while long lists were 80 words long. A test list following a short study list comprised all 20 words that appeared on the study list as well as 20 distractors. Using the retroactive design, test lists following long study lists included the first 20 words studied and 20 distractors. In all test lists, half of both the targets and distractors were high frequency and half were low frequency. The words in the test lists were presented in a random order.

Each study-test list combination was randomly assigned a different font color (blue, black, red and green) for each subject. All words were presented in lower-case letters in the center of the computer screen on a grey background.

During study, each word appeared for 3000ms and within that time subjects were also required to rate its pleasantness on a six point Likert scale (1: least pleasant, 6: most pleasant) by clicking on the appropriate button displayed across the computer screen below the probe word. Subjects were instructed that if they missed rating the pleasantness of any word within the allocated time, they were to move on and rate the next word.

Subjects were given a three second warning before the onset of each test list. Responses in the test lists took the form of the yes/no recognition paradigm. Words were presented individually along with two possible response options "old" and "new" which appeared as buttons on the computer screen. Subjects were informed that they were to respond "old" if they recognized the word from the preceding study list and to respond "new" if they believed that to be the first presentation of the word by clicking on the appropriate button. Words appeared on screen until such time as the subject responded.

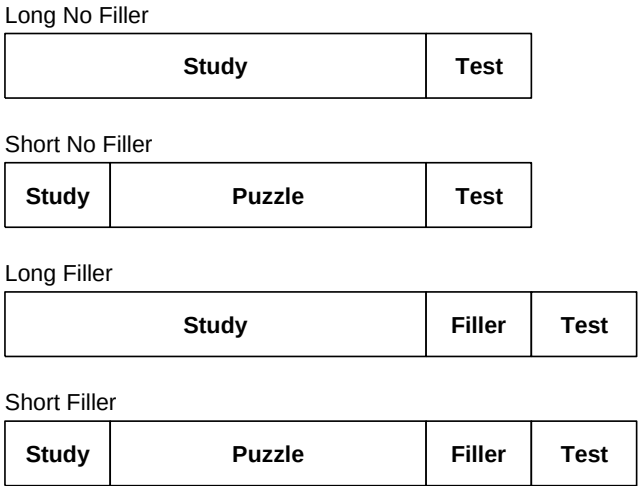

Figure 1. The design of the recognition memory experiment. Puzzle activity was added to equate retention interval.

To equate the retention interval for the short and long lists, all short lists were followed by three minutes of a sliding tile puzzle activity. Thus the total time elapsed between the first word of the short list and the test list was equal to the time between the presentation of the first word on the long list and the subsequent test list. To encourage contextual reinstatement, this experiment included a condition in which both the short and long lists were followed by eight minutes of sliding tile puzzle filler in addition to the three minutes following the short list (see Figure 1).

The experiment was counterbalanced for order. Half of the subjects began with the filler condition while half did not. Within each of these conditions, half of the subjects began with the short list and the remainder started by viewing a long list. To ensure continuity, both the short and long lists from within the one condition were studied one after the other, before the subject studied the lists from the other condition. All subjects took part in each condition and there were no missing data.

Results

The yes/no recognition procedure provides two independent counts per subject per condition: A hit count and a false alarm count. Test items that appear on the study list are called targets and test items that did not are called distractors. The hit count is the number of target items to which the subject responded yes. The false alarm count is the number of distractors to which the subject erroneously responded yes. For a given number of targets and distractors, these two counts determine correct rejection and miss counts.

Using these counts, it is straightforward to apply standard Signal Detection Theory (e.g., Macmillan & Creelman, 1991) to model recognition memory. The model is shown in Figure 2. The key assumption is that the evidence the test item appeared on the study list lies on a uni-dimensional strength continuum[1]. Recognition strengths are drawn from two separate distributions, one corresponding to target words and the other corresponding to distractor words. The distributions are assumed to be Gaussian and have equal variance (we will discuss unequal variance versions of the model later in the paper), but the mean strength is higher for the target words. Decisions are made by comparing the recognition strength to a fixed criterion, denoted by *k*, and choosing 'yes' for those words above criterion, and 'no' for those words below criterion. As shown in Figure 2, these stimulus and decision-making assumptions correspond to predictions about hit, false-alarm, miss, and correct-rejection rates that can be related to the experimental data.

The main benefit of the Signal Detection Theory model is that it provides separate measures of discriminability and bias. Discriminability is a measure of how distinct target words are from distractor words, and so corresponds to how well people perform on the yes/no task. Bias measures to what extent they more inclined to give 'yes' or 'no' responses, regardless of their level of performance. There are a number of ways discriminability and bias can be measured, which are all just reparameterizations according to the model in Figure 2. In this paper, we use the 'd-prime' measure of discriminability, denoted, *d*, which is the distance between the means of the target and distractor distributions[2]. We also use the *c* measure of bias, which is the signed difference between the criterion *k* and the unbiased criterion value at which false alarms and misses are equally likely. Larger values of *d* correspond to better performance on the task. Positive values of *c* correspond to a bias towards saying 'no', and so produce higher miss rates. Negative values of *c* correspond to a bias towards saying 'yes', and so produce higher false-alarm rates.

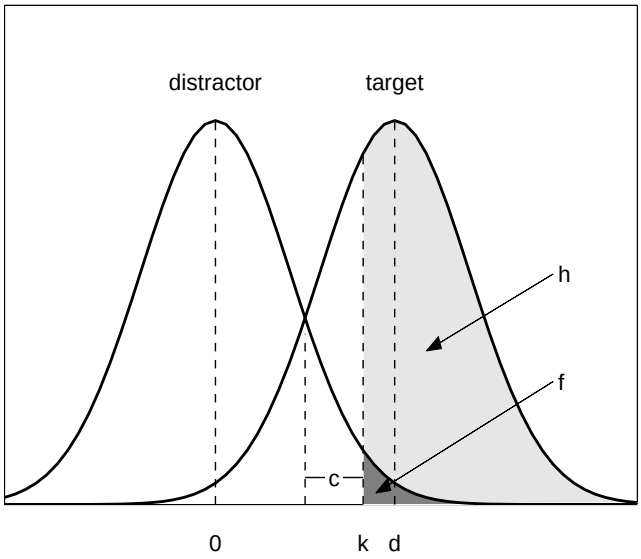

Figure 2. The Signal Detection Theory model of recognition memory.

We undertook a standard analysis to estimate these measures of discriminability and bias. This involved, first, deriving hit and false alarm rates for each subject by dividing their hit and false alarm counts by, respectively, the number of targets and distractors. These hit rates, *H*, and false alarm rates, *F*, were then used to calculate *d* and *c* val-

[1] The uni-dimensional assumption is not that only one source of information contributes to the decision. Rather, recognition memory models that use SDT typically assume that there are a very large number of sources of evidence that are relevant (Murdock, 1982; Hintzman, 1984; Humphreys et al., 1989). However, the assumption is that these sources are condensed to a single value in order to arrive at a decision.

[2] Since only the differences between the distributions is important, the distractor distribution is usually given a mean of zero, and the target distribution a mean of *d*.

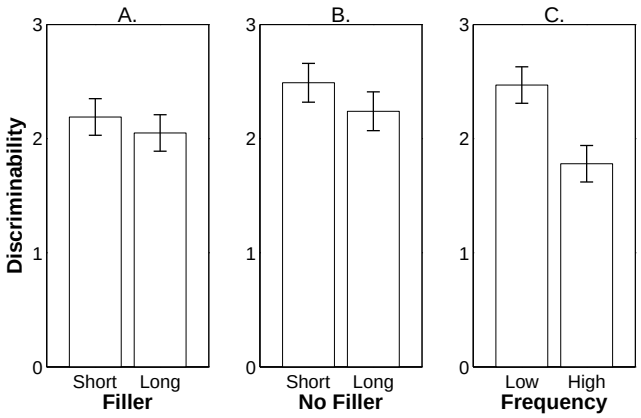

Figure 3. Means and 95% confidence intervals for discriminability in the (A) filler, (B) no filler and (C) word frequency comparison.

ues, according to the formulae (e.g., Macmillan & Creelman, 1991)

$$d = z(f) - z(h), \quad (1)$$

$$c = \frac{z(h) + z(f)}{2}. \quad (2)$$

Hit rates of 1.0 or false alarm rates of 0.0 imply an infinite value for the *d* measure of discriminability. To overcome this problem, we followed the advice of Snodgrass and Corwin (1988), and added 0.5 to the hit and false-alarm counts and 1 to the target and distractor counts.

Figure 3 shows the means and 95% confidence intervals for discriminability in the filler, no filler, and word frequency comparisons. In the filler comparison, where contextual reinstatement was encouraged after both the short and long lists, a repeated measures ANOVA yielded a nonsignificant effect of list length on *d* ($F(1, 47) = 1.65$, $p = .21$). Conversely, in the no filler condition, where the contextual reinstatement control was relaxed, a statistically significant effect of list length on *d* was found ($F(1, 47) = 4.44$, $p = .04$, $\eta_p^2 = .09$), suggesting that list length did have an effect on performance. In the word frequency comparison, a statistically significant effect on *d* was found ($F(1, 47) = 117.98$, $p < .001$, $\eta_p^2 = .72$), with low frequency words being better discriminated than high frequency words.

These results indicate no list length effect when filler activity was employed, but a list length effect when the filler activity was removed. In addition, the results indicate, consistent with established findings, that low frequency words were more easily discriminated than high frequency words. Thus, according to this standard analysis, one would conclude that contextual reinstatement can induce a list length effect and a failure to control for this confound will lead to artifactual list length findings.

Six Inferential Problems

In this section, we discuss six problems with the standard analysis. Problems 1–3 relate to model selection, and are caused by deficiencies in the NHST approach to inference. Problems 4–6 relate to parameter estimation, and are caused by deficiencies in the frequentist approach to estimation.

Problem 1: Evidence in Favor of the Null. The NHST approach to inference seeks to establish if there is sufficient evidence to suggest that the mean for target words is different from the mean for distractor words. This is inappropriate when both the null and alternative hypotheses have theoretical weight. NHST assumes that the null hypothesis is true until the data prove otherwise. In practice, issues of potentially low power ensure that only significant effects favoring the alternative hypothesis are considered theoretically useful. This makes it impossible to find evidence for the theoretical position that predicts the absence of a list length effect. What is needed are models that can directly assess the evidence in favor of any theoretical position.

Problem 2: Iterative Use. NHST cannot be applied in an iterative way, where current results are examined before deciding whether to collect additional data. Because NHST does not conform to the likelihood principle, the sample size must be fixed before running the experiment (Wagenmakers, in press). This is constraining in cases where results approach significance, and it is possible only a few extra subjects would have been required. It is also wasteful in cases where the effect turns out to be much larger than expected, and it is necessary to continue experimentation until the planned sample size is reached, particularly in the

context of research with special populations. What is needed are models that permit iterative testing.

Problem 3: Inference from the Majority. NHST attempts to establish if there is a difference between two means, without regard to the proportion of subjects contributing to that difference. If enough subjects are tested, a difference is certain to be found, no matter how small the proportion of subjects exhibiting an effect. If there are individual differences in recognition memory, a minority of subjects may evidence an effect, and the standard analysis will infer a general property of the memory system from these subjects. What is needed are models that are not unduly influenced by a minority of subjects. A related concern is the method for excluding subjects from analysis, for which current practices vary widely. What is needed are models that are not overly sensitive to exclusion decisions.

Problem 4: Small Sample Sizes. NHST makes assumptions about its sampling distributions that rely on asymptotic results. This means it is not necessarily valid with small sample sizes. What is needed are models that are guaranteed to be valid for any sample size.

Problem 5: Edge Corrections. As explained above, it is common for frequentist estimators of hit and false-alarm rates to imply infinite measures of d. These estimates require an *ad hoc* edge correction, but the correction chosen can have a large effect on the results. What is needed are models that do not require edge corrections.

Problem 6: Capturing Sampling Variability. The frequentist estimators of hit and false-alarm rates fail to account for the uncertainty in these rates that must exist given finite data. If a subject has 3 hits from 4 targets, their hit rate is much less certain than if they have 30 hits from 40 targets. The standard analysis is insensitive to the number of data from which hit and false-alarm rates are estimated. What is needed are models that are sensitive to uncertainty about hit and false-alarm rates.

Our primary concern in this paper is with the analysis procedures that are commonly used in the memory literature and the potentially problematic influence they may be having on our understanding of recognition, in particular. However, several of the issues that we outline above are well known in the broader statistics literature and several branches of frequentist statistics have been developed to address these problems. Equivalency tests provide a means of testing if two means are approximately equal (Wellek, 2003), stopping rules adjust required significance values to avoid taking advantage of chance if one discontinues an experiment upon consideration of the already collected data (e.g., Armitage, 1958), robust statistics provide values to replace the mean that are less subject to distortion by outliers (e.g., Hampel, Ronchetti, Rousseeuw, & Stahel, 1986), inference from the majority is possible using mixture models estimated using the expectation maximization algorithm (Greun & Leisch, 2006), and a variety of exact tests can be applied when sample sizes are small (e.g., Fisher, 1922). These approaches are, however, rarely employed in the recognition literature. In the next section, we develop a Bayesian method which we believe is more elegant and more straightforward to apply.

A Bayesian Approach

In this section, we develop a Bayesian approach that overcomes the problems with the standard analysis outlined above. [3] We focus first on the parameter estimation problem of inferring discriminability and bias measures from experimental data, and then move to the model selection problem of comparing competing list length and no list length accounts.

Parameter Estimation

Bayesian inference represents what is known and unknown about the variables of interest using probability distributions. These distributions provide complete representations of uncertainty, and automatically solve the parameter estimation problems 4–6. That is, using the Bayesian approach means that the measures of discriminability inferred from data take into account sampling variability, never need edge corrections, and are valid for any sample size.

[3] Matlab and WinBUGS code for the method is available at http://www.socsci.uci.edu/ ~mdlee/

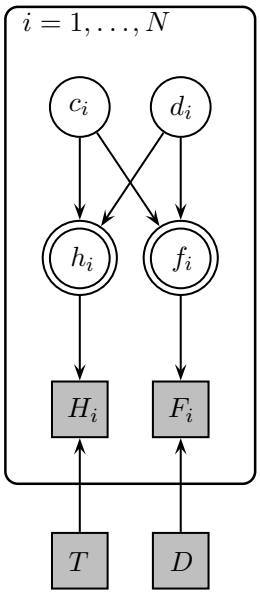

Figure 4. Graphical model for inferring discriminability and bias from hit and false-alarm counts in a yes/no recognition memory experiment.

Graphical models provide a convenient formalism for expressing many Bayesian models (e.g., Jordan, 2004). Worked examples of this approach to psychological modeling can be found in Lee (2008, in press); Lee and Wagenmakers (2008) and Shiffrin, Lee, Wagenmakers, and Kim (submitted). The basic idea is that the model is represented by a directed graph, with nodes corresponding to variables, and the dependencies between variables captured by edges, with each child node depending on its parents. We use the conventions that observed variables have shaded nodes, while unobserved variables are not shaded, and continuous variables have circular nodes while discrete variables have square nodes. We also use plates to denote independent replication of parts of the graph. In addition, where it aids interpretation, and introduces a meaningful psychological variable, we introduce deterministic unobserved variables, shown as double-bordered nodes.

We use the graphical model shown in Figure 4 to infer measures of discriminability d_i and bias c_i for the *i*th subject. To estimate these measures, we use the SDT model to reparameterize discriminability and bias into a hit rate h_i and a false-alarm rate f_i of the *i*th subject, according to the relationship

$$h_i = \Phi(\frac{1}{2}d_i - c_i), \quad (3)$$

$$f_i = \Phi(-\frac{1}{2}d_i - c_i). \quad (4)$$

In the graphical model in Figure 4 this reparameterization is shown by the d_i, c_i, h_i and f_i nodes. The d_i and c_i nodes are shown as unshaded circles to indicate they are both continuous variables with unknown values. The h_i and f_i nodes are also circular, because they are also continuous, but have double-borders, because they follow deterministically from their parent d_i and c_i nodes.

The model places priors on discriminability and bias that correspond to the assumption of uniform priors for the hit and false-alarm rates, as follows

$$d_i \sim \text{Gaussian}(0, 2), \quad (5)$$

$$c_i \sim \text{Gaussian}(0, \frac{1}{2}). \quad (6)$$

There are four counts for each condition for each subject that constitute their observed data. These are all shown as square and shaded nodes, because they take discrete values and are observed. The number of target trials, T, and the number of distractor trials, D, are the same for all subjects in our experiment, and so are placed outside the plate. The hit count, H_i and the false-alarm count F_i vary across subjects. We assume the hit and false-alarm counts follow a Binomial distribution depending on the hit and false-alarm rates, and the number of target and distractor trials, so that

$$H_i \sim \text{Binomial}(T, h_i), \quad (7)$$

$$F_i \sim \text{Binomial}(D, f_i). \quad (8)$$

It is straightforward to implement the model in Figure 4 using free WinBUGS software (Lunn, Thomas, Best, & Spiegelhalter, 2000), which provides the capability to sample from the posterior (i.e., the distributions of the variables conditional on the observed data) using standard Markov-Chain Monte-Carlo computational methods (see Chen, Shao, & Ibrahim, 2000; Gilks, Richardson, & Spiegelhalter, 1996; Mackay, 2003).

Figure 5 shows the posterior distributions for discriminability, hit rate and false-alarm in three

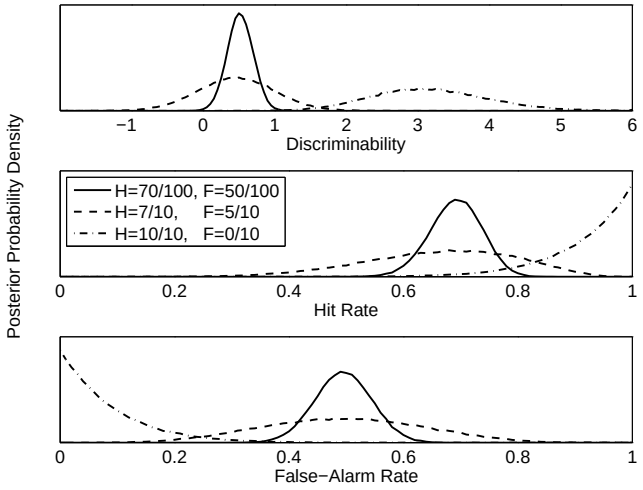

Figure 5. Posterior distributions for discriminability, hit rate and false-alarm rate in three illustrative situations.

illustrative situations, based on 10^4 posterior samples. In the first situation, 70 hits and 50 false-alarms are observed in 100 target and 100 distractor trials. Because of the large number of trials, there is relatively little uncertainty surrounding the hit and false-alarm rates, with narrow posteriors centered on 0.7 and 0.5 respectively. Discriminability is known with some certainty, centered on about 0.5. In the second situation, 7 hits and 5 false-alarms are observed in 10 target and 10 distractor trials. These are the same rates of hit and false-alarms of the first situation, but based on many fewer samples. Accordingly, the posterior distributions have (essentially) the same means, but show much greater uncertainty. In the third situation, perfect performance is observed, with 10 hits and no false-alarms in 10 target and 10 distractor trials. The modal hit and false-alarm rates are 1.0 and 0.0, but other possibilities have some density, and so discriminability is well defined.

Taken together, these illustrations show how the Bayesian approach solves the estimation problems 4–6. Comparing the first and second situation shows how posterior distributions are sensitive to the uncertainty inherent in sampling variability. The third situation shows that using the full distribution avoids the need for edge corrections. And posterior distributions can validly be found in exactly the same way using any sample size.

Model Selection

In a Bayesian analysis, competing theoretical positions are represented by models, which can be compared directly to each other based on data. In all of our experimental comparisons, the main theoretical question is whether there is a systematic change in discriminability between two experimental conditions, measured subject by subject according to the within-subjects design. For the filler and no filler conditions, the interest is in whether short lists have better discriminability than long lists. For the word frequency comparison, the interest is in whether low frequency words are more discriminable than high frequency words.

Two Competing Models. For all of these comparisons, we consider two competing models. The 'Error-Only' model assumes the within-subject differences in discriminability come from a Gaussian distribution of unknown variance, but with a mean of zero. This model captures the assumption that there is no systematic difference in discriminability, although there will inevitably be noisy variation in the differences. Formally, the difference in discriminability between the first and second conditions for the ith subject, $\Delta d_i = d_i^A - d_i^B$ is modeled as

$$\Delta e_i \sim \text{Gaussian}\,(0, \lambda_e). \tag{9}$$

The alternative 'Error-plus-Effect' model assumes the within-subject differences in discriminability follows the sum of a Gamma distribution and a zero-mean Gaussian distribution. This corresponds to the idea that there is a systematic positive difference, as well as the noisy variation. Formally, Δd_i is modeled as

$$\Delta f_i = f_i^e + f_i^f. \tag{10}$$

where f_i^f is an effect component drawn from a Gamma distribution,

$$f_i^f \sim \text{Gamma}\,(\alpha, \beta), \tag{11}$$

and f_i^e is an error component again drawn from a zero-mean Gaussian distribution

$$f_i^e \sim \text{Gaussian}(0, \lambda_e). \quad (12)$$

The competing accounts of how each subjects' discriminability changes across the different list length conditions can be seen visually in the way d_i^A relates to d_i^B in the graphical model. These two discriminabilities differ by Δd_i for the ith subject, and this change is itself either generated by the error-only account (on the left side, involving Δe_i) or by the error-plus-effect account (on the right side, involving Δf_i).

We assume standard near non-informative prior on the variances for the error components (see Spiegelhalter, Thomas, & Best, 1996)

$$\lambda_e \sim \text{InverseGamma}(0.001, 0.001), \quad (13)$$
$$\lambda_f \sim \text{InverseGamma}(0.001, 0.001); \quad (14)$$

and follow George, Makov, and Smith (1993) in placing a near non-informative prior on the parameters of the effect component

$$\alpha \sim \text{Exponential}(1), \quad (15)$$
$$\beta \sim \text{Gamma}(0.1, 0.1). \quad (16)$$

These priors on the variance and hyper-parameters of the effect distribution are the least 'principled' ones in our model (i.e., their basic forms have a good justification, but their exact values are more arbitrary). In our analyses, we explored a large number of variations, such as using InverseGamma $(0.01, 0.01)$ for the variances, or a Gamma $(0.01, 0.01)$ prior on the β hyper-parameter in the effect distribution. These changes all had only a slight quantitative effect on the results, and supported exactly the same conclusions.

Mixture Model Comparison. One standard Bayesian method for comparing models is to calculate the Bayes Factor, which measures how much more likely the data are to have arisen under one model rather than the other. The Bayes Factor, however, potentially does not meet our requirement of basing its inference on the behavior of the majority of subjects. Because of the all-or-none loss function the Bayes Factor seeks to optimize (i.e., the Bayes Factor assumes one model is correct and the other is incorrect), it is possible for one or a few extreme subjects to over-ride the evidence of the majority. For example, if almost all subjects behave according to a simpler model, but one or two behave according to a much more complicated model, the Bayes Factor will report evidence in favor of the more complicated model. This is because, under these circumstances, if only one model is true, and has to explain the behavior of all subjects, that model has to be the more complicated one.

As a more satisfactory alternative, following Lee (2008), we use a Bayesian approach to model selection based on mixture estimation. The key idea is that, rather assuming exactly one model is correct and the other is incorrect, we assume both models are useful, but one may be more useful (i.e., be more likely to explain the behavior of more subjects) than the other. For our current problem, this approach means making inferences for each subject as to whether they are better modeled by the error-only account or by the effect-and-error account, and using this information to make an inference about underlying *rate* at which subjects are assigned to the two accounts. The behavior of a small number of subjects can have only a limited effect of the overall rate of assignment, and so the conclusions are robust. And, as with all fully Bayesian methods of model selection, the inference process is automatically sensitive both to goodness-of-fit and model complexity.

Combining the two models and their mixture comparison gives the graphical model show in Figure 6. The binary indicator variable x_i controls which account is used to model the difference Δd_i for the ith subject,

$$\Delta d_i = \begin{cases} \Delta e_i & \text{if } x_i \text{ is } 0 \\ \Delta f_i & \text{if } x_i \text{ is } 1. \end{cases} \quad (17)$$

and each x_i has probability θ of selecting the error-plus-effect account

$$x \sim \text{Bernoulli}(\theta), \quad (18)$$

where we use assume a flat prior for the rate θ

$$\theta \sim \text{Uniform}(0, 1). \quad (19)$$

This approach means that each subject is conceived as following either the error-only or error-plus-effect account, as specified by their x_i binary variable, and that there is a fixed proportion of subjects who do each, as specified by the θ rate variable. From finite data, however, there will always be uncertainty about both which account each subject belongs to, and the exact value of the underlying rate of belonging. This means there will be a posterior distribution for the x_i variables (i.e., each subject will have some number of posterior samples in the error-only account and some number in the error-plus-effect account), as well as a posterior distribution over the rate θ. The uncertainty about belonging in the posteriors for the x_i variables is naturally represented in the posterior of the θ mixing proportion, which essentially measures the probability any subject will be classified as an error-only versus error-plus-effect subject.

Bayesian Results

We again implemented the graphical model in Figure 6 in WinBUGS. We obtained 5×10^4 posterior samples for the three theoretical comparisons—list length with filler, list length without filler and word frequency—after a burn-in period of 10^3 samples (i.e., a set of samples that are discarded, to allow the MCMC sampling to adapt to the stage where it is sampling from the posterior distribution), and using multiple chains (i.e., multiple independent runs of the sampling process with different starting points) to diagnose convergence.

Overall Rate of Assignment. Figure 7 presents the posterior distributions of the rate θ at which subjects are best modeled by the error-plus-effect model for each comparison. The most likely rates are small for the list length comparisons, indicating that most subjects are better modeled by the error-only account. In contrast, for the word frequency comparison, the most likely rates are large, indicating that subjects are most likely better modeled by the error-plus-effect account. These results show how the Bayesian approach solves model selection problem 1, because it is possible to find evidence directly for the 'null' error-only model, as well as for the 'alternative' error-plus-effect model.

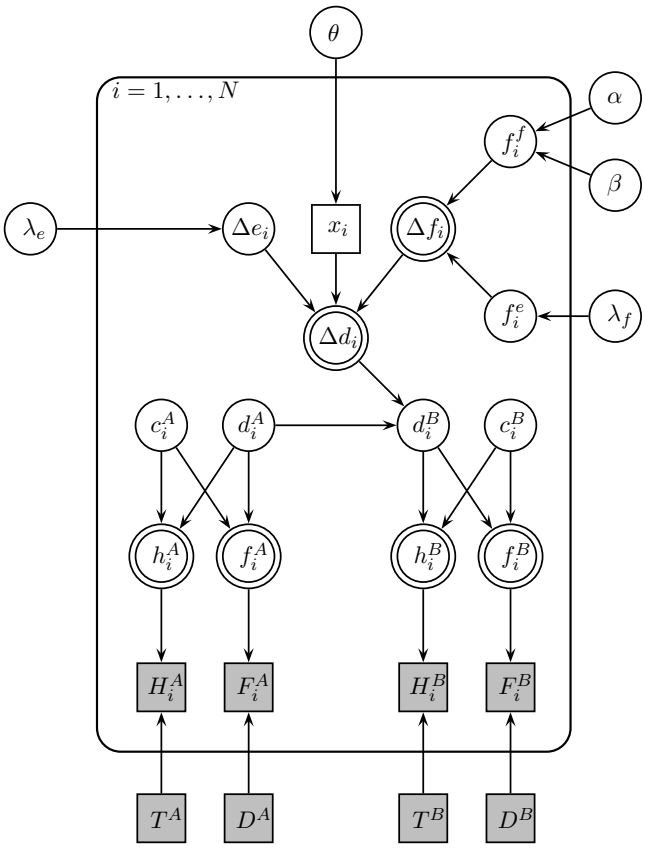

Figure 6. Graphical model for inferring the rate subjects belong to the error-plus-effect versus error-only accounts of the change in their discriminability between experimental conditions.

Individual Subject Analysis. More detail on the Bayesian model results is provided by Figure 8 which shows, in the upper panels, the posterior predictive distributions of the error-only and error-plus-effect accounts for all three comparisons. These correspond to the expected distribution of differences in discriminability under the two competing models, based on the experimental data. Figure 8 also shows, in the lower panels, the modeled mean and 95% credible intervals for the observed differences in discriminability for each subject. Those subjects most often assigned to the error-only account have means shown by white circles, while those most often assigned to the error-plus-effect account have means shows by black circles. Note that, even though the most likely as-

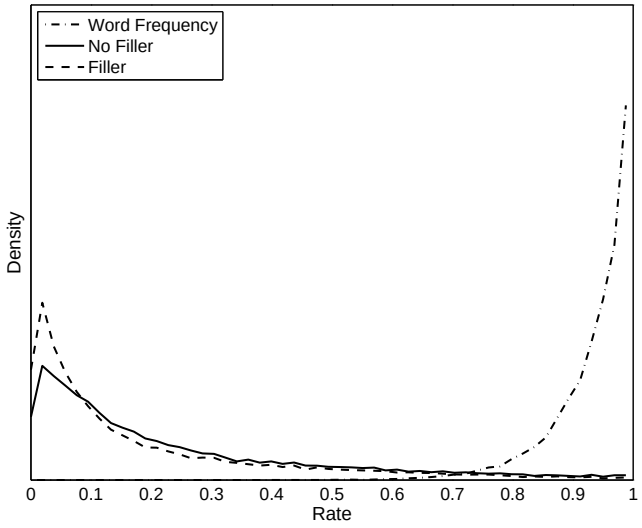

Figure 7. Posterior distribution of the underlying rate subjects are better modeled by the error-plus-effect account for the filler, no filler, and word frequency comparisons.

signment for each subject is the same in each analysis (always error-plus-effect for word frequency, but always error-only for the two list length conditions), these assignments are not certain, and the posterior distribution over the θ rates in Figure 7 represents this uncertainty.

Iterative Analysis. A basic property of the Bayesian approach is that inferences can be made at any stage of data collection, and so the method can be applied iteratively. To demonstrate this, we found the posterior rates θ for the first $6, 12, \ldots, 48$ subjects in each comparison. As a summary measure of each posterior distribution, we then calculated the proportion of the rate posterior between 0 and 0.1, between 0.1 and 0.9, and between 0.9 and 1.0. The idea is that these three categories correspond to support for just the error-only model, for both models, and for just the error-plus-effect models, respectively. In this way, we can summarize the full posterior distribution of θ by three proportions that sum to one, and tell us whether one or the other, or both, competing models are useful in explaining the data.

The results of this analysis are shown in Figure 9. Panels A, B, and C correspond the word frequency, no filler, and filler comparisons, respectively. In each panel, the three possibilities are represented as the vertices of a triangle, and the relative weight given to each as the iterative analysis progresses is shown by a path in this triangle. The shading inside the triangle corresponds to critical proportions of 0.5, 0.7 and 0.9 in favor of each possibility. It is clear that the word frequency comparison quickly provides strong evidence for the error-plus-effect account, while both the no filler and filler comparisons provide strong evidence for the error-only account. This demonstration makes it clear that the Bayesian approach solves model selection problem 2. In an iterative experiment, it would be statistically justified to terminate data collection once a pre-determined critical level was reached, and so collect data from as many subjects as required to reach a conclusion.

Panel D of Figure 9 shows the proportions for the no filler comparison resulting from excluding the one, two, three or four most extreme subjects favoring the error-plus-effect account from the analysis, as well as for the original full data analysis. Because the points are nearby, the various exclusions do not greatly affect the conclusions that would be drawn. In general, estimating the rate of assignment will not change drastically if a few subjects are excluded. This is why the Bayesian method makes inference based on majority behavior, and so addresses model selection problem 3. In contrast, we note that starting with the full data set and then excluding the same one, two, three or four subjects from the no filler data under NHST analysis, the F values decrease from 4.44 to 2.81 to 2.02 to 1.29 to 0.83, with an associated increase in the p-values from 0.04 to 0.10 to 0.16 to 0.26 to 0.37. In this case, the exclusion of a single subject is sufficient to change the substantive conclusion.

Extension to Unequal Variance SDT

One reasonable criticism of the preceding analysis is that its use of the equal-variance form of SDT may be inappropriate. As reviewed by Mickes, Wixted, and Wais (2007), empirical examination of Receiver Operating Characteristic (ROC) curves in SDT analyses of recognition memory data typ-

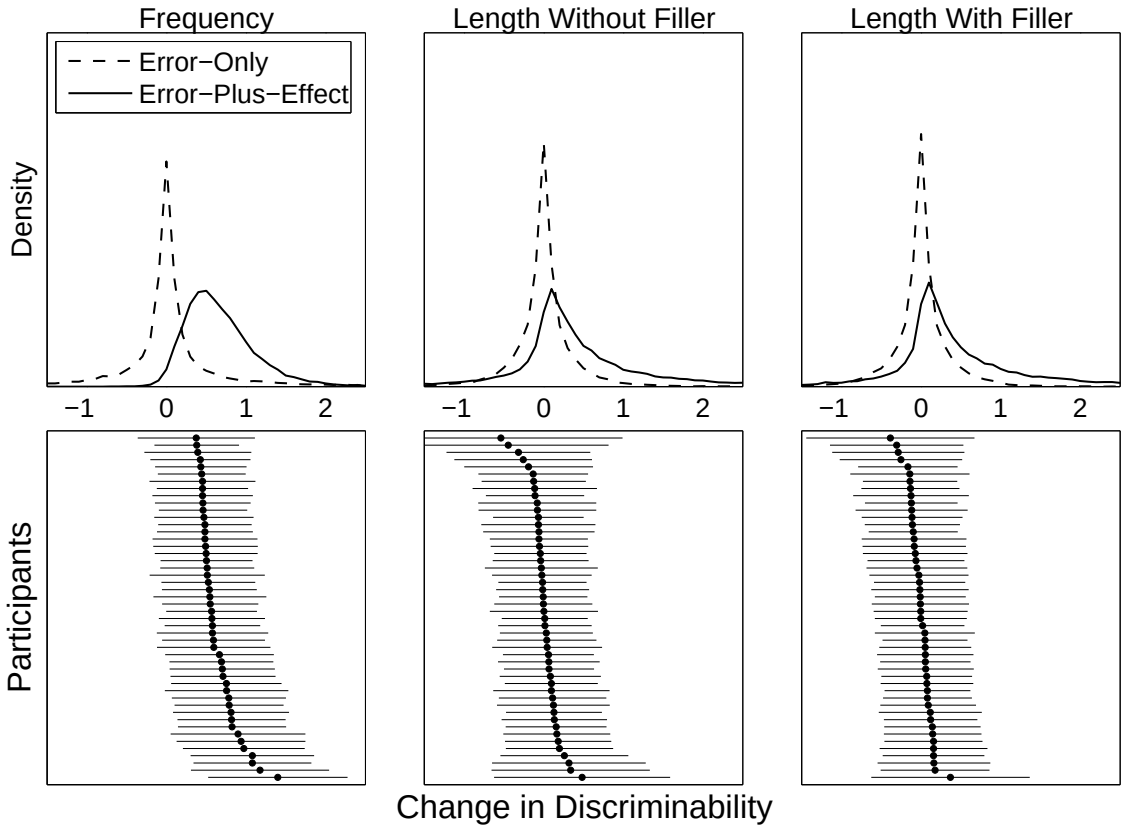

Figure 8. How the error-only and error-plus-effect accounts model the change in subjects discriminability between conditions for the (A) word frequency, (B) list length without filler, and (C) list length with filler comparisons. The upper panels show how each account models the distribution of differences in discriminability. The lower panels show the mean and 95% credible intervals for the change in discriminability for each subject. Means shown in white or black correspond to subjects most often assigned to the error-only or error-plus-effect account, respectively.

ically show that the standard deviation of the distractor distribution, σ_d, is smaller than the standard deviation of the target distribution, σ_t. Mickes et al. (2007) present additional supporting empirical evidence for this difference, and argue that the standard deviation of the distractor distribution is approximately only 80% as large as for the target distribution, so that $\sigma_d/\sigma_t \approx 0.8$.

A straightforward way to extend our analysis to allow for unequal-variances within the SDT framework is to incorporate direct assumptions about the ratio of standard deviations. In the graphical model this means introducing a new observed variable $\tau = \sigma_d/\sigma_t$, which specifies an assumed level for unequal variances, and redefining the false-alarm rates to be

$$f_i = \Phi(-\frac{1}{\tau}(d_i/2 - c_i)). \quad (20)$$

All other part of the model can remain unchanged. [4] This extended model includes the original equal-variance model as a special case when $\tau = 1$, but allows assumptions such as $\tau = 0.8$, or any other value of interest, to be examined.

Figure 10 shows the results of a series of analyses of the posterior rate of assignment making

[4] We note that there is an unavoidable theoretical indeterminacy in the unequal-variance SDT model, because the distractor and target distributions intersect twice. We follow previous practice and avoid this problem by taking the practical stance that it is the point of intersection between the means of the distribution that is of psychological interest.

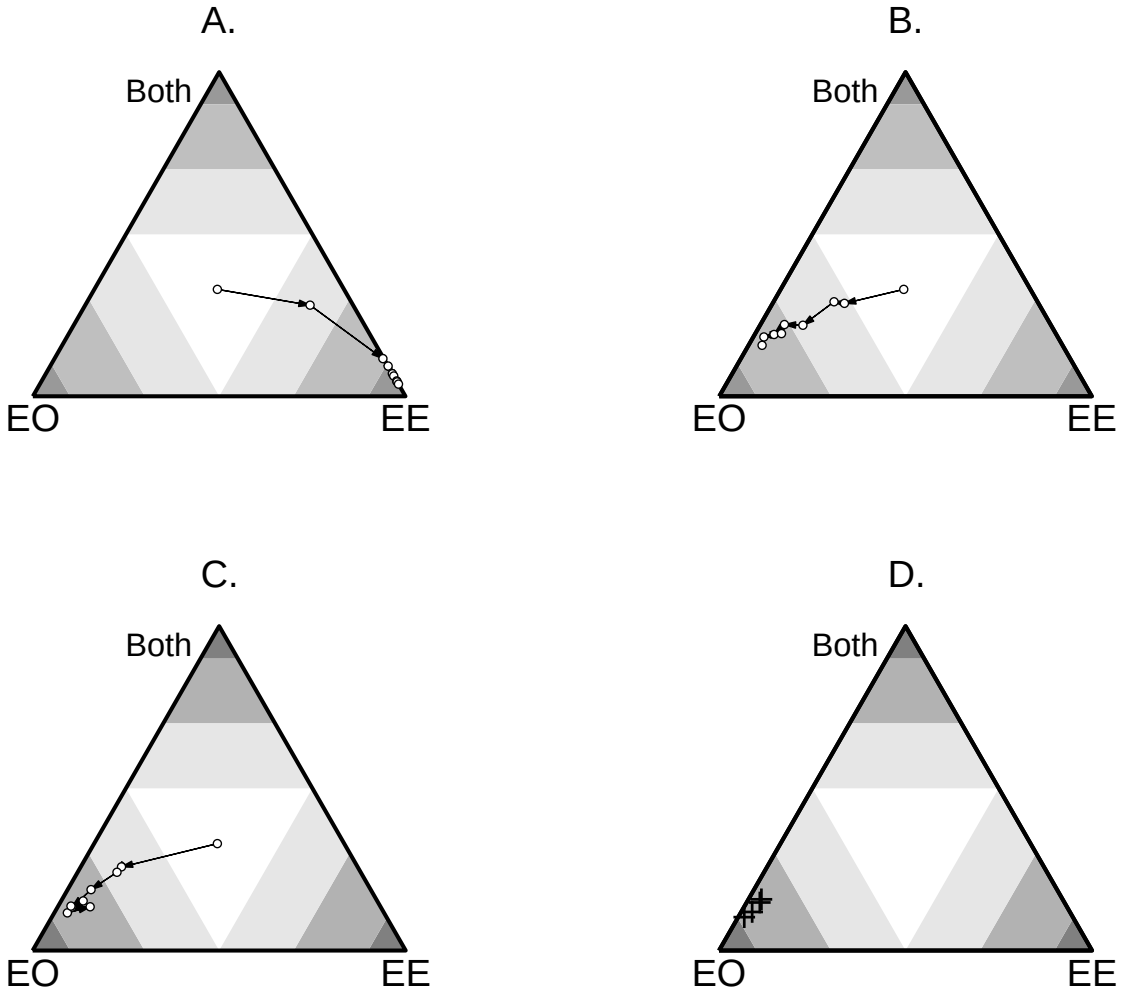

Figure 9. Iterative analyses for the (A) word frequency, (B) list length without filler, (C) list length with filler, and (D) exclusion of extreme subjects in the list length without filler comparisons. Each panel shows the error-only (EO), error-plus-effect (EE) and both possibilities, corresponding to proportions of the rate posterior θ between 0 and 0.1, 0.1 and 0.9, and 0.9 and 1.0, respectively. For Panels A–C, the paths show these proportions in iterative analyses, adding another 6 subjects on each iteration. For Panel D, the crosses show the proportions resulting from excluding the one, two, three or four most extreme subjects in favor of the error-plus-effect account from the list length without filler analysis, as well as for the original full analysis.

different assumptions about the ratio of variances. The analyses use τ = 0.1, 0.5, 0.8, 1, 2, and 10, and so consider cases where the distractor distribution is both more and less variable than the target distribution. It is clear that the same conclusions found in the equal-variance case in Figure 7 hold for all but the most extreme assumptions about unequal variances (i.e., when τ = 10 and the distractor distribution has a ten-fold larger standard deviation than the target distribution). For the 'reasonable' assumptions about τ, where the distractor distribution is less variable, especially including the τ = 0.8 advocated by Mickes et al. (2007), there remains strong evidence for an effect in the word-frequency condition, but strong evidence for no effect in the list-length conditions.

Discussion

Recognition memory involves bringing together information about the test item and its context

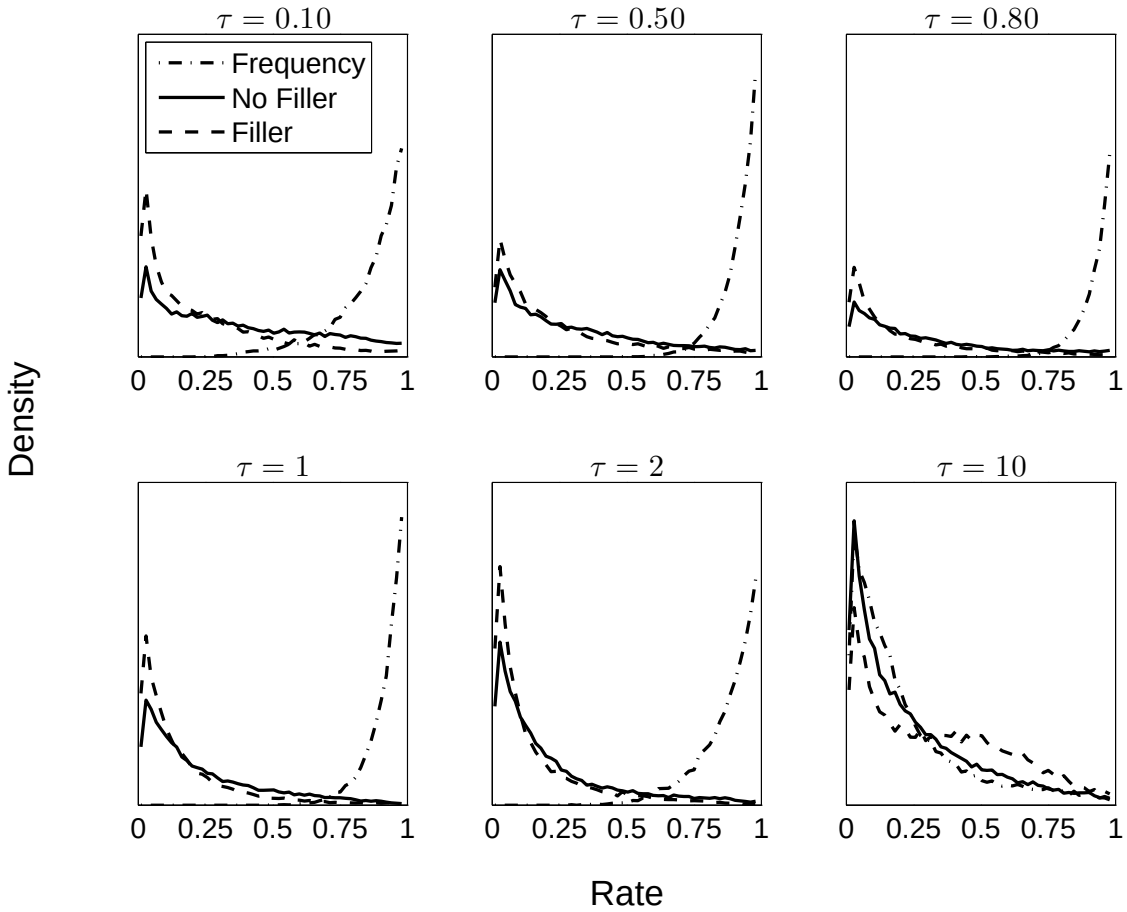

Figure 10. Posterior distribution of the underlying rate subjects are better modeled by the error-plus-effect account for the filler, no filler, and word frequency comparisons for six different assumptions about the ratio of standard deviations for the distractor and target distributions.

(Humphreys et al., 1994). Consequently, interference in the paradigm can logically derive from one or both of two sources. Item noise models propose that interference comes primarily from the other items that appeared in the study list, while context noise models propose that interference comes primarily from the other contexts in which the test item has been seen.

The majority of existing models of recognition assume an item noise approach (Gillund & Shiffrin, 1984; Murdock, 1982; Eich, 1982; Hintzman, 1986; Humphreys et al., 1989; Shiffrin & Steyvers, 1997; McClelland & Chappell, 1998; Clark & Gronlund, 1996). However, Dennis and Humphreys (2001) argued that by proposing that recognition is a context noise process, while recall is an item noise process, one can make sense of several key dissociations between the procedures.

First, performance on low frequency words is better than for high frequency words in recognition, whereas either no effect or a high frequency advantage is typically found in recall (Glanzer & Adams, 1985; Gillund & Shiffrin, 1984). If recognition is dominated by context noise, low frequency words will be subject to less interference, and will be recognized better. If recall is not subject to context noise, low frequency words will be subject to the same level of interference, and no effect will be found.

Second, if subjects are presented with lists constructed from weak and strong items—where strength is manipulated either by study duration or number of presentations—recognition performance on same strength items is the same as if

subjects are presented with lists constructed purely of weak items, or purely of strong items (Ratcliff et al., 1990). This phenomenon is called the null list-strength effect. In recall, however, strengthening some items in a list impairs performance on the unstrengthened items. If recognition is dominated by context noise, the strengthening of other items will not impact performance. If recall is dominated by item noise, strengthening other items increases interference for unstrengthened items.

The final, and most controversial, line of argument involves the list length effect. List length has an agreed substantial effect in recall, but a debated effect in recognition. Dennis and Humphreys (2001) argued that a number of potential confounds including retention interval, attention, rehearsal and contextual reinstatement could lead to artifactual list length effects. When they controlled for these confounds Dennis and Humphreys (2001) found no list length effects. However, Cary and Reder (2003) have contested this conclusion finding a list length effect using similar controls.

The status of the list length effect is particularly important in distinguishing between models of recognition memory because item noise models have been developed that can account for the word frequency and null list strength results. The prediction of a list length effect, however, would seem to be an inescapable consequence of the item noise assumption.

Furthermore, because of its unresolved status, the presence or absence of the list length effect makes an ideal case study for improving the analysis of recognition memory experiments. The current standard frequentist methods for estimation, and null hypothesis significance testing methods for model selection, have a number of undesirable properties. Using NHST, evidence cannot be found in favor of the possibility there is no list length effect. The results of NHST can be determined by a small proportion of subjects, contrary to the aim of inferring general properties of the memory system. NHST requires a fixed sample size be established before experimentation begins. These sample sizes must be large for the statistical assumptions of NHST to be sound, and sufficiently large sample sizes are guaranteed to reject the null hypothesis. Frequentist point estimates of discriminability are insensitive to the uncertainty associated with sampling variability, and require edge corrections that can have a large effect on the results.

In this paper, we have developed and applied a Bayesian approach to understanding recognition memory using Signal Detection Theory. We demonstrated how the Bayesian method overcomes the problems with the standard methods by applying both to a new set of data. The fact that the Bayesian analysis found evidence for the absence of a list length effect for words supports a context noise account of recognition memory. Of course, this result relates to one data set only. Thus, the primary source of interference in recognition memory remains an open question. But, we believe our development of powerful Bayesian methods for inference and analysis of recognition memory experiments is a crucial step towards reaching an answer.

References

Armitage, P. (1958). Sequential methods in clinical trials. *American Journal of Public Health Nations Health*, *48*(10), 1395-1402.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81-99.

Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory and Cognition*, *11*, 307-315.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*(2), 231-248.

Chen, M. H., Shao, Q. M., & Ibrahim, J. G. (2000). *Monte carlo methods in bayesian computation*. New York: Spinger-Verlag.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and Review*, *3*(1), 37-60.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452-478.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, *89*(6), 627-661.

Fisher, R. A. (1922). On the interpretation of 2 from

contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, *85*(1), 87-94.

George, E. I., Makov, U. E., & Smith, A. F. M. (1993). Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, *20*(2), 147-156.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain monte carlo in practice*. Boca Raton (FL): Chapman & Hall/CRC.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1-67.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition*, *13*(1), 8-20.

Greun, B., & Leisch, F. (2006, April). Finite mixture model diagnostics using the parametric bootstrap. In W. Elmenreich & H. Kaiser (Eds.), *Proceedings of the junior scientist conference 2006* (p. 301-302).

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1335-1369.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Jphn Wiley & Sons.

Hintzman, D. L. (1984). Minerva-2 - a simulation-model of human-memory. *Behavior Research Methods Instruments & Computers*, *16*(2), 96-101.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*(4), 411-428.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system - a theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208-233.

Humphreys, M. S., Wiles, J., & Dennis, S. (1994). Toward a theory of human-memory - data-structures and access processes. *Behavioral and Brain Sciences*, *17*(4), 655-667.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140-155.

Lee, M. D. (2008). Three case studies in the bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1-15.

Lee, M. D. (in press). BayesSDT: Software for bayesian inference with signal detection theory. *Behavior Research Methods*.

Lee, M. D., & Wagenmakers, E. J. (2008). *A course in bayesian graphical modeling for cognitive science.* (Unpublished lecture notes, University of California, Irvine. `http://www.socsci.uci.edu/~mdlee/CourseBook_v1.pdf`)

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100-109.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

Mackay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724-760.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*, 858-865.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609-626.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology-Learning Memory and Cognition*, *17*(5), 855-874.

Ohrt, D. D., & Gronlund, S. D. (1999). List length effect and continuous memory: Confounds and solutions. In C. Izawa (Ed.), *On human memory: Evolution, progress and reflections on the 30th anniversary of the atkinson shiffrin model* (p. 105-126). Mahwah, NJ: Erlbaum.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect .1. data and discussion. *Journal of Experimental Psychology-Learning Memory and Cognition*, *16*(2), 163-178.

Schulman, A. L. (1974). The declining course of recognition memory. *Memory and Cognition*, *2*, 14-18.

Shiffrin, R. M., Lee, M. D., Wagenmakers, E. J., & Kim, W. J. (submitted). A survey of model evaluation approaches with a focus on hierarchical bayesian methods.

Shiffrin, R. M., & Steyvers, M. (1997). Model for recognition memory: Rem - retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145-166.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34-50.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1996). *Bugs examples volume 1, version 0.5.* Cambridge, UK.

Underwood, B. J. (1978). Recognition memory as a function of the length of study list. *Bulletin of the Psychonomic Society*, *12*, 89-91.

Wagenmakers, E. J. (in press). A practical solution to the pervasive problem of p-values. *Psychonomic Bulletin and Review*.

Wellek, S. (2003). *Testing statisical hypotheses of equivalence*. Chapman & Hall/CRC.