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

Simon Dennis
School of Psychology
University of Adelaide

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

Angela Kinnell
School of Psychology
University of Adelaide

Abstract

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 critical test of competing item- and context-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 different models representing different theoretical assumptions about memory.

In a typical yes/no recognition memory task, participants 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 (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, changes in the length of the study list should not change recognition performance. There is no consensus on whether or not a list length effect is observed empirically. Dennis and Humphreys (2001) argued that, for verbal stimuli, context is the primary source of interference, 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 using a remember-know paradigm.

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 standard method of analysis 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 how memory works.

Experiment

Participants

Forty-eight Psychology students from the University of Adelaide participated in the study. There were 10 males and 38 females with ages ranging from 16 to 42

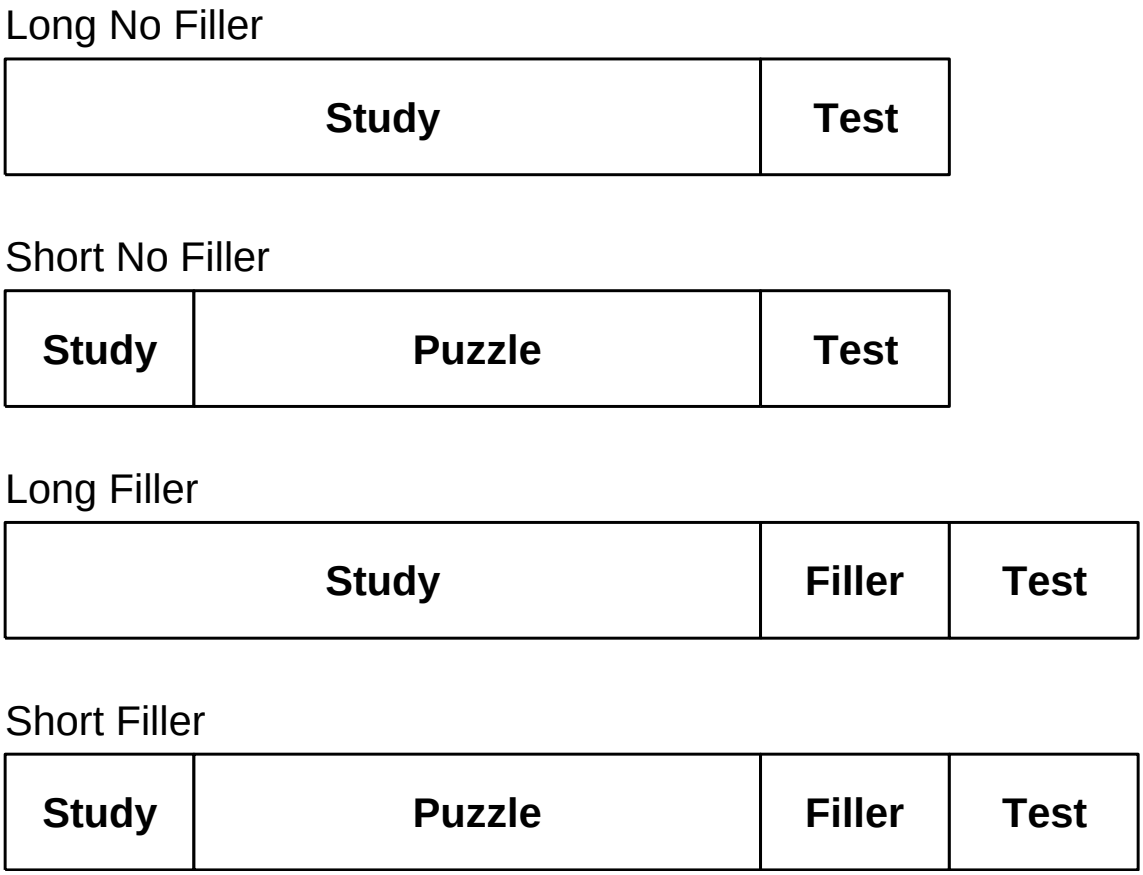

Figure 1. The design of the recognition memory experiment. Puzzle activity was added to equate retention interval.

years ($M = 19.71$, $SD = 4.73$). The sample size was equivalent to that in Dennis and Humphreys (2001), and larger than in the Cary and Reder (2003) study.

Method

The four conditions of our experiment are presented in Figure 1. The study sets were either long, with 80 words, or short, with 20 words, and participants were required to make pleasantness ratings of each word during study. Each set of words contained both low frequency and high frequency words. Test sets contained 20 words that were always chosen from the first 20 words of a long list. Short study lists were always followed by puzzle activity equating the time between study and test with that of long lists. Finally, extra 'filler' conditions, involving an additional activity before test, were used for both the long and short lists.

This experimental design controls for four confounding variables, each capable of artifactually producing a list length effect, that are inherent in the standard design where testing immediately follows studying either a short or long list (Dennis & Humphreys, 2001). First, the retention interval between study and test is controlled by the retroactive design, the puzzle activity, and selecting test words from the first 20 words in long lists. Second, the retroactive design and requirement to make pleasantness ratings encourages participants to attend to all of the words in long lists as they do for short lists. Third, a demanding sliding puzzle was used to prevent participants rehearsing words between study and test for short lists. Finally, the filler versus no filler manipulation measures the effect of contextual reinstatement. Because short lists had puzzle activity that long lists did not, testing on short lists demanded the study context be reinstated, but testing on long lists could be done maintaining the current study context. Since test words were always chosen from the beginning of long lists, the final study context might differ substantially from when target words were studied, thus compromising performance on the long list. The filler conditions control for this potential confound through their inclusion of additional unrelated activity.

Results

The yes/no recognition procedure provides two independent counts per participant 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 participant responded yes. The false alarm count is the number of distractors to which the participant 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, 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.

[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). The assumption is that these sources are condensed to a single value.

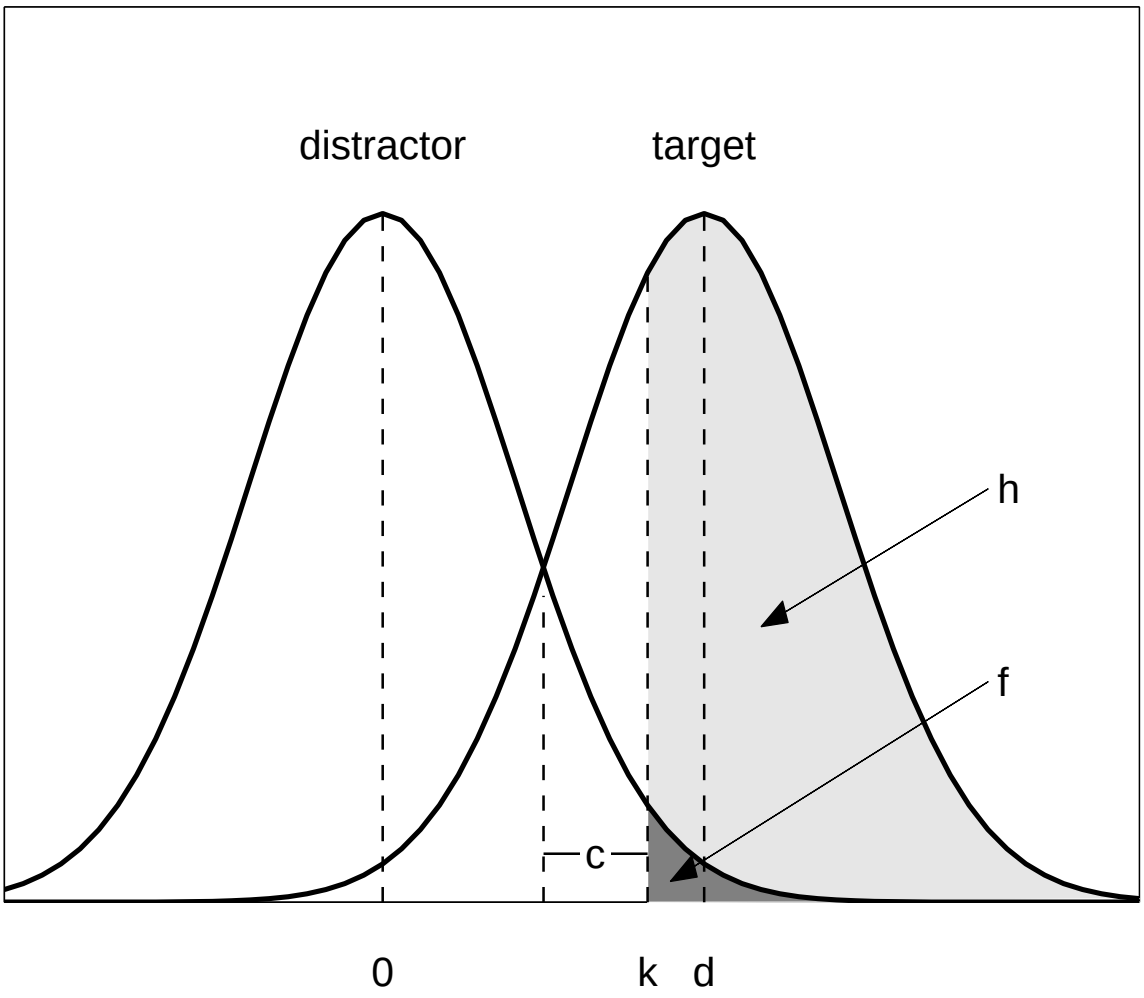

Figure 2. The Signal Detection Theory model of recognition memory.

The main benefit of the Signal Detection Theory model is that it provides distinct 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

[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.

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.

We undertook a standard analysis to estimate these measures of discriminability and bias. This involved, first, deriving hit and false alarm rates for each participant 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 values, according to the formulae (e.g., Macmillan & Creelman, 1991)

$$d = z(f) - z(h), \tag{1}$$

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

A common problem with these calculation is 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

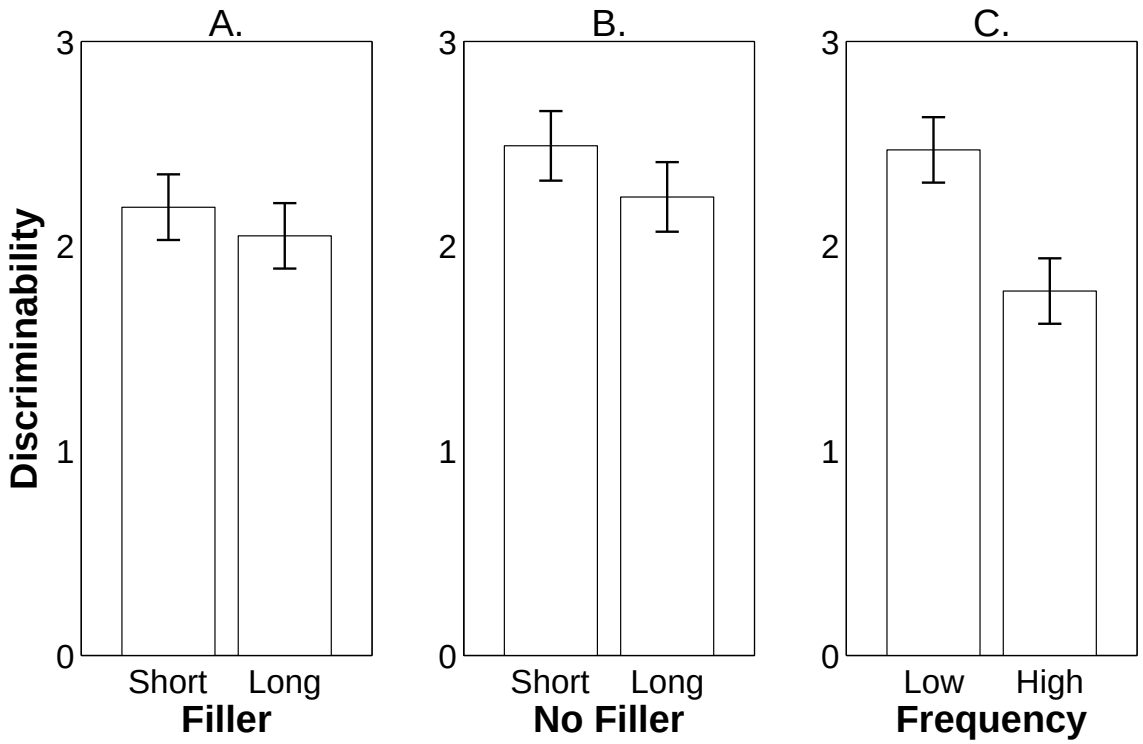

Figure 3. Means and 95% confidence intervals for discriminability in the (A) filler, (B) no filler and (C) word frequency comparison.

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 decided 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 participants 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 the two means, without regard to the proportion of participants contributing to that difference. If enough participants are tested, a difference is certain to be found, no matter how small the proportion of participants exhibiting an effect. If there are individual differences in recognition memory, a minority of participants may evidence an effect, and the standard analysis will infer a general property of the memory system from these participants. What is needed are models that are not unduly influenced by a minority of participants. A related concern is the method for excluding participants 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 fails to account for the uncertainty in these rates that must exist given finite data. If a participant 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.

A Bayesian Approach

In this section, we develop a Bayesian approach that overcomes the problems with the standard analysis. 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.

Graphical models provide a convenient formalism for expressing many Bayesian models (e.g., Jordan, 2004). 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 the part of the graph inside the bounding box. In addition, where it aids interpretation, 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 ith participant. 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 ith participant, according to the relationship

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

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

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

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

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

There are four counts for each condition for each participant that constitute their observed data. The number of target trials, T, and the number of distractor trials, D, are the same for all participants in our experiment, and so are placed outside the plate. The hit count, H_i and the false-alarm count F_i vary across participants. 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

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

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

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

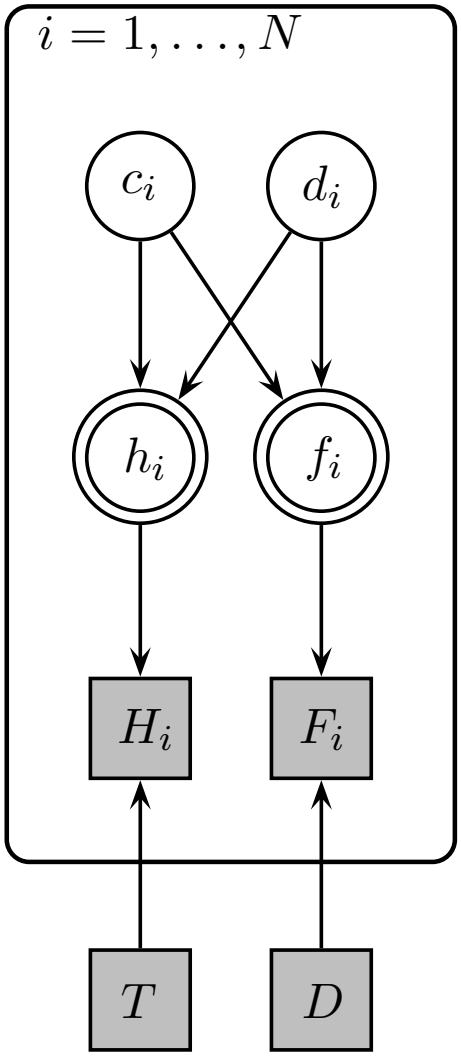

Figure 4. Graphical model for inferring discriminability and bias from hit and false-alarm counts in a yes/no recognition memory experiment.

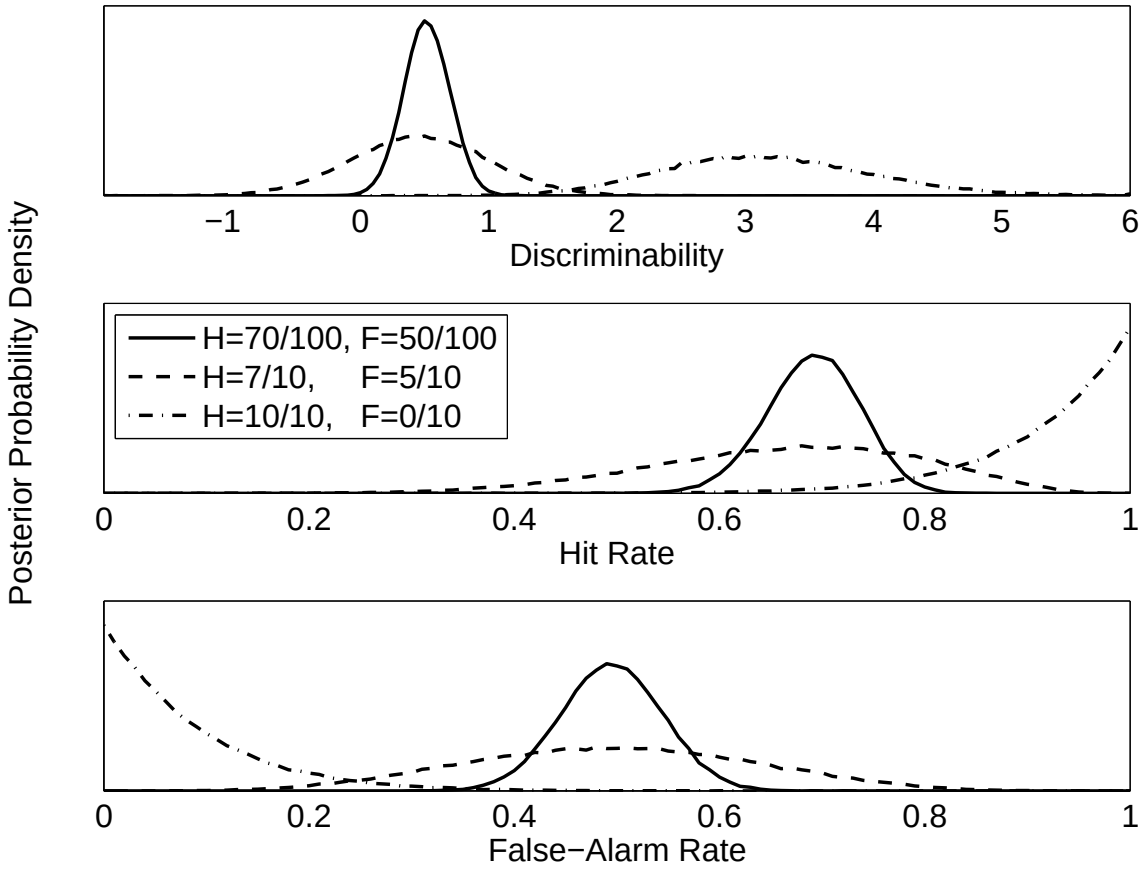

Figure 5. Posterior distributions for discriminability, hit rate and false-alarm rate in three illustrative situations.

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 samping 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 participant by participant 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 participant, $\Delta d_i = d_i^A - d_i^B$ 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

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

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

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

and follow George, Makov, and Smith (1993) in placing a vague prior on the parameters of the effect component

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

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 participants. Because of the all-or-none loss function the Bayes Factor seeks to optimize, it is possible for one or a few extreme participants to over-ride the evidence of the majority. As a more satisfactory alternative, we use an alternative Bayesian approach to model selection based on mixture estimation. The key idea is that inferences are made for each participant as to whether they are better modeled by the error-only account or by the effect-and-error account, and the underlying *rate* at which participants are assigned to the two models is used as the comparative measure. The behavior of a small number of participants can only have 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 discriminability and bias for each participant in both conditions are found independently, and the error-only and error-plus-effect accounts are then compared as models of differences in discriminability. The binary indicator variable x_i controls which account is used to model the difference Δd_i for the ith subject,

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

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

$$x \;\sim\; \text{Bernoulli } (\theta). \tag{18}$$

Finally, we use assume a flat prior for the rate θ

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

Bayesian Results

It is straightforward to implement the model in Figure 6 using free WinBUGS software (Spiegelhalter, Thomas, & Best, 2004), 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). We obtained 2×10^4 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, and using multiple chains to diagnose convergence.

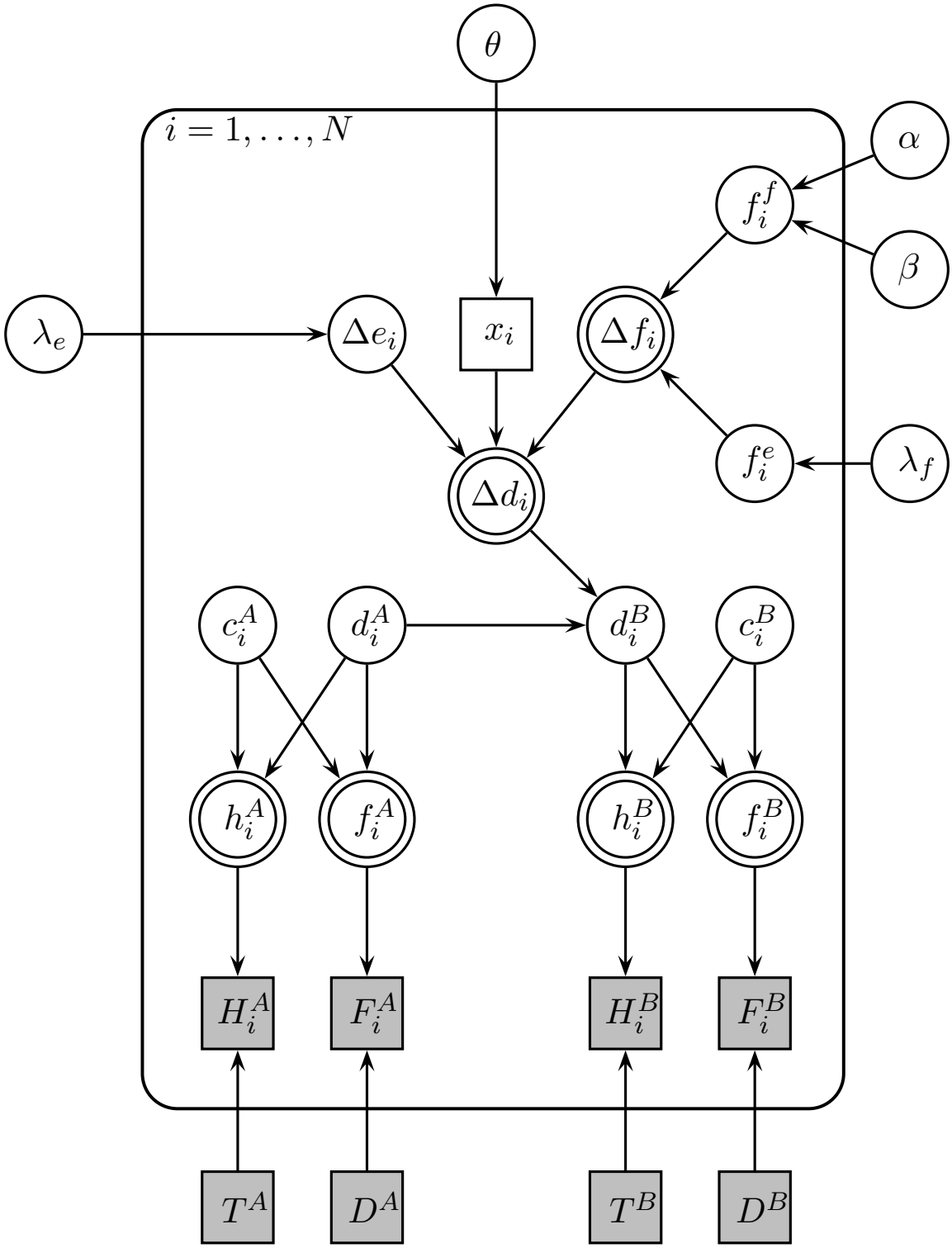

Figure 6. Graphical model for inferring the rate participants belong to the error-plus-effect versus error-only accounts of the change in their discriminability between experimental conditions.

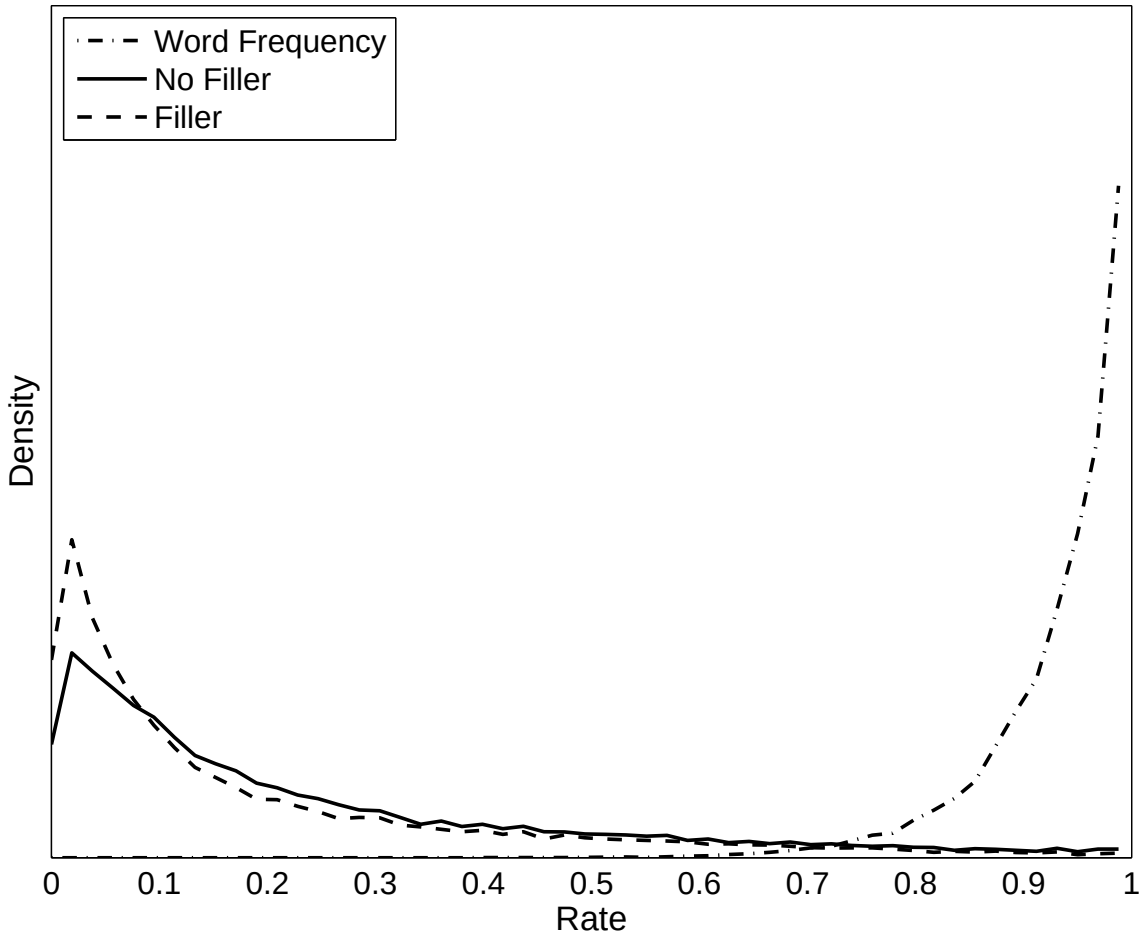

Figure 7. Posterior distribution of the underlying rate participants are better modeled by the error-plus-effect account for the filler, no filler, and word frequency comparisons.

Figure 7 presents the posterior distributions of the rate θ at which participants 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 participants are better modeled by the error-only account. In contrast, for the word frequency comparison, the most likely rates are large, indicating that most participants are 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.

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 participants. Those participants 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.

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 $6, 12, \ldots, 48$ participants 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 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 participants 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 participants 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 participants 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 by excluding the same one, two, three or four participants from the no filler data under NHST analysis, the F values decrease from 2.81 to 2.02 to 1.29 to 0.83, with an associated increase in the p-values from 0.10 to 0.16 to 0.26 to 0.37.

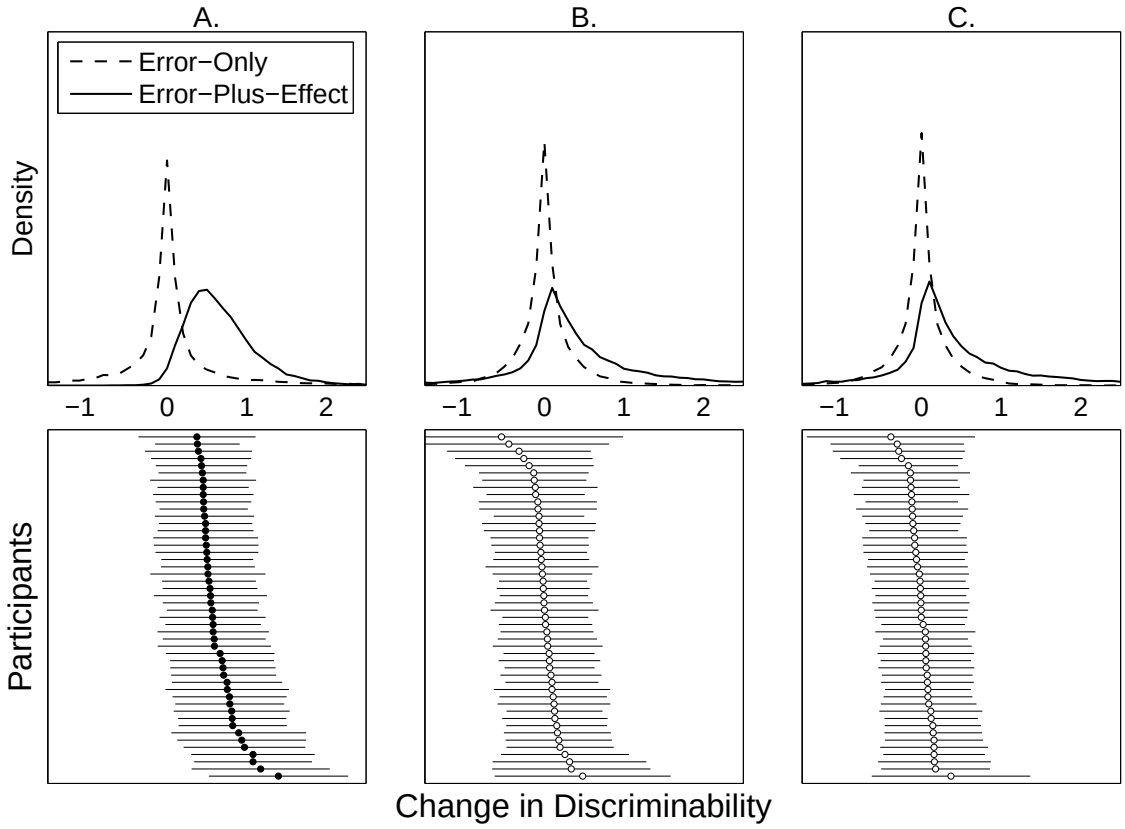

Figure 8. How the error-only and error-plus-effect accounts model the change in participants discriminability between conditions for the (A) word frequency, (B) list length without filler, and (C) list length with filler comparisons. The upper panels show how each account models the distribution of differences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

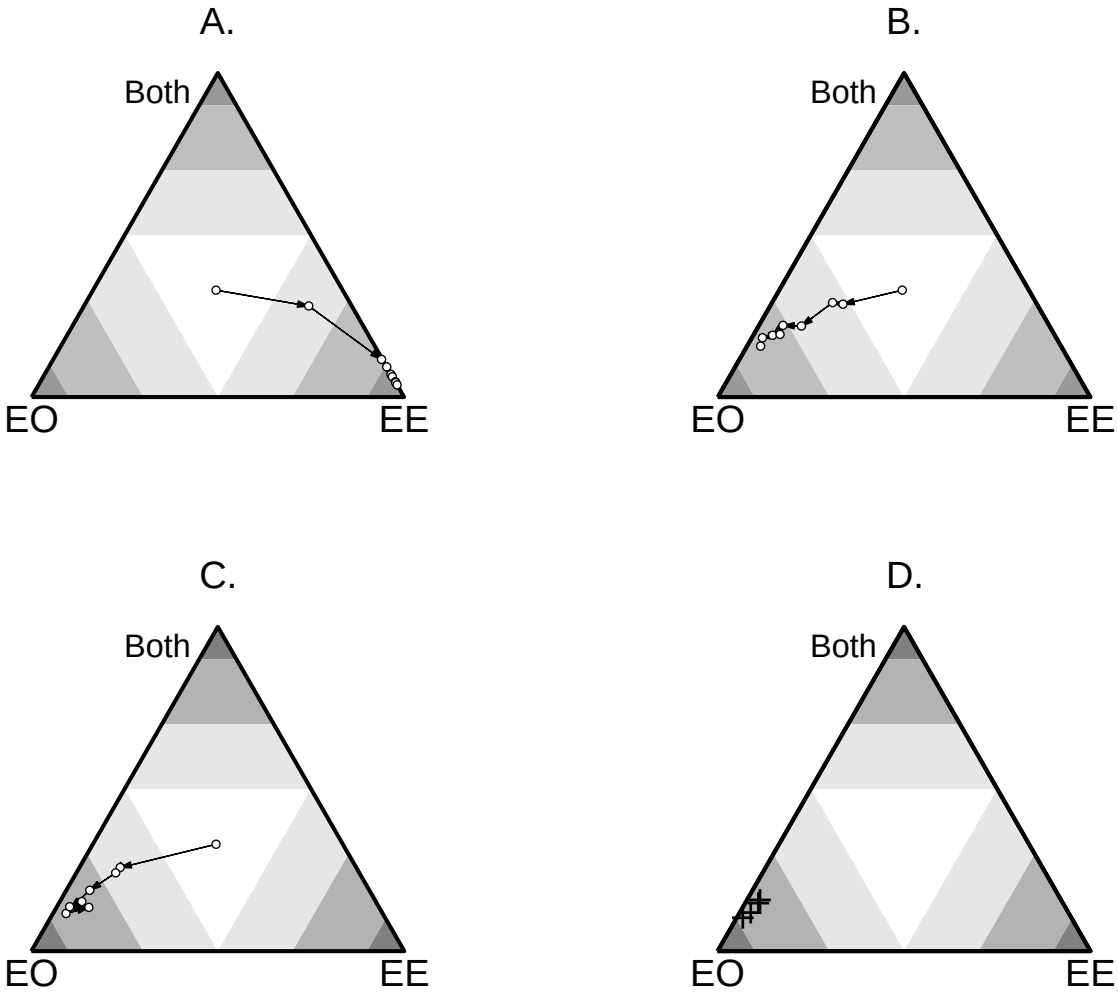

Figure 9. Iterative analyses for the (A) word frequency, (B) list length without filler, (C) list length with filler, and (D) exclusion of extreme participants in the list length without filler comparisons. Each panel shows the error-only (EO), error-plus-effect (EE) and both possibilities, corresponding 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 participants on each iteration. For Panel D, the crosses show the proportions resulting from excluding the one, two, three or four most extreme participants in favor of the error-plus-effect account from the list length without filler analysis, as well as for the original full analysis.

Discussion

Recognition memory involves bringing together information about the test item and its context (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. 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 participants are presented with lists constructed from weak and strong items—where strength is manipulated either by study duration or number of presentations—recognition performance is the same as if participants 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.

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 participants, 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, and the list length effect is not the only relevant empirical test of competing item and context theories of interference. 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

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

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.

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 & Cognition*, *13*(1), 8-20.

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., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*(3), 662–668.

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

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

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

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

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. (2004). *WinBUGS version 1.4 user manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS examples volume 1, version 0.5*. Cambridge, UK: MRC Biostatistics Unit.

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