Journal of Memory and Language 85 (2015) 27–41

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The list-length effect does not discriminate between models of recognition memory

Jeffrey Annis [a,c], Joshua Guy Lenes [a], Holly A. Westfall [a], Amy H. Criss [b], Kenneth J. Malmberg [a,*]

[a] University of South Florida, United States
[b] Syracuse University, United States
[c] Vanderbilt University, United States

ARTICLE INFO

Article history:
Received 13 February 2015
revision received 31 May 2015

Keywords:
Recognition memory
List-length effect
Memory models
Bayesian
BCDMEM
REM

ABSTRACT

Dennis, Lee, and Kinnell (2008) claimed that they obtained evidence for a null list-length effect (LLE) for recognition memory, and that their finding was consistent with context-noise models and inconsistent with item-noise models of memory. This claim has since been repeated in several articles (e.g., Kinnell & Dennis, 2011; Turner, Dennis, & Van Zandt, 2013). However, a more thorough investigation of their data indicates that Dennis et al.'s findings are inconclusive, and their assertion that empirical observations of the LLE may distinguish between item-noise and context-noise models is debatable. In fact, their findings provide very little evidence in favor of a null LLE; there is actually a credible positive LLE in one condition of their experiment, a finding that context-noise models cannot explain. Moreover, we show that Dennis et al.'s findings support an item-noise model like the retrieving effectively from memory (REM) at least as well as a context-noise model. The source of the erroneous conclusions is identified as the measurement model Dennis et al. developed. In the end, we conclude that the list-length effect obtained from present experimental designs is insufficient for competitively testing item-noise and context-noise models of recognition.

© 2015 Published by Elsevier Inc.

A long standing issue in the memory literature concerns whether increasing the number of items studied during an episodic memory experiment affects memory accuracy when tested via a recognition procedure. The phenomenon is known as the list-length effect (LLE), and Dennis, Lee, and Kinnell's (2008) goal was to determine if changes in list length are consistent with context-noise models like BCDMEM, which predicts a null LLE (Dennis & Humphreys, 2001) or item-noise models like REM, which predicts a small, positive LLE (Criss & Shiffrin, 2004a, 2004b; Criss, Malmberg, & Shiffrin, 2011; Shiffrin & Steyvers, 1997). The contribution of their research hinged on the assertion that, "The status of the list-length effect is particularly important in distinguishing between models of recognition memory …the prediction of a list-length effect, however, would seem to be an inescapable consequence of the item noise assumption" (Dennis et al., 2008, p. 372). Based on their new experiment, they concluded that increasing the number of items studied does not affect recognition memory and that their findings are consistent with context-noise models and inconsistent with item-noise models.

A more thorough investigation of their findings indicates that Dennis et al.'s assertion that empirical observations of the LLE may distinguish between item-noise and context-noise models is debatable. We begin by introducing the theoretical issue and present several new analyses of Dennis et al.'s data that indicate straightforward conclusions based on the observed LLE are not supported. Then we describe the motivations for the Bayesian analysis

[*] Corresponding author at: Dept of Psychology, USF, PCD4118G, 4202 East Fowler Ave., Tampa, FL 33611, United States.

E-mail address: malmberg@usf.edu (K.J. Malmberg).

http://dx.doi.org/10.1016/j.jml.2015.06.001
0749-596X/© 2015 Published by Elsevier Inc.

conducted by Dennis et al., how it was implemented, and what they concluded from it. We show that their strong conclusions are not supported by their data and that their conclusions are based on invalid tests of the models in question. As such, we conclude that the ability of the LLE to discriminate between models of recognition memory that predict a null LLE and a small LLE is doubtful.

Some history

For many years, researchers believed that increasing the number of items studied produces decreases in recognition accuracy. This was partly due to early reports of an LLE with recognition memory testing, partly due to similar findings in tests of memory using recall procedures, and partly due to intuition; "everyone knows it is more difficult to remember many as compared to a few items" (see Gillund & Shiffrin, 1984, for a review). However, there were some reports of null LLEs in the literature, especially when several confounding nuisance variables were controlled (e.g. Koppell, 1977). One problem for establishing a null LLE is that the frequentist statistical analyses did not provide any basis for support of a null effect, and it is still unknown whether the "controls" that were implemented were not confounded with a different set of nuisances, as we will see below when we discuss the design of the experiment conducted by Dennis et al. (2008). Nevertheless, the issue of whether there was a null or small LLE became theoretically important with the development of several new models of recognition memory. Dennis and Humphreys (2001), in particular, proposed a "context-noise" model in which recognition is based on the match between the context used to probe memory and the context associated with the prior occurrence of the test item, and only the test item. Their model, known as BCDMEM, broke with traditional models in several ways, and it predicts a null LLE.

The difference between BCDMEM and other models of recognition memory is that BCDMEM assumes that the episodic representation of an encounter with a word on the study list only consists of the context information in which the encounter occurs. These context features are shown in the left panel of Fig. 1 and they are labeled "Retrieved Context Features". The retrieved context features consist of the context features stored during study in addition to the context features stored during all prior encounters with the word. Variability in the features representing the context of different encounters with a word is what defines different episodic traces. A critical assumption of BCDMEM is that information about the word itself is not encoded during any event; the representation of the word exists prior to the experiment and it is associated with new context features. The word representations are labeled "Item Units" in Fig. 1.

At test, when a word stimulus is presented for a recognition judgment in BCDMEM, only the corresponding item unit is activated and this instigates the retrieval of the context features previously associated with the item unit. The null LLE is predicted based on this assumption because only the representation of the test stimulus takes part in the retrieval process. The number of items studied is irrelevant to the particular contexts in which the word has occurred.

Some of the retrieved context features correspond to the context features present during the encoding of words that were studied (targets) and some of the retrieved context features correspond to prior encounters with the word. If the word was not studied (foil), then the retrieved context features are those associated with the word previously. The retrieved context features are compared to context features that are "mentally reinstated" by the subject and correspond to the context features present during study. The reinstatement of context features occurs independently of the context features retrieved as the result of the item unit activation. The stimulus is judged to be a member of the study list if the match between the reinstated context features and the retrieved context features exceeds a decision criterion. Since targets were recently associated with the reinstated context, they are more likely to be positively endorsed than foils.

Other models, like the retrieving effectively from memory (REM) model, assume that during an encounter with a word on a study list, information is stored about the word, such as its orthography, phonology, and meaning, and information is stored about the context in which the word was encountered. This is shown in the right panel of Fig. 1 where a longer list is illustrated and a shorter list is illustrated. Each episodic memory trace consists of both item and context features.

At test, a retrieval cue consisting of item features representing the word stimulus and context features is compared to the contents of memory; the greater the

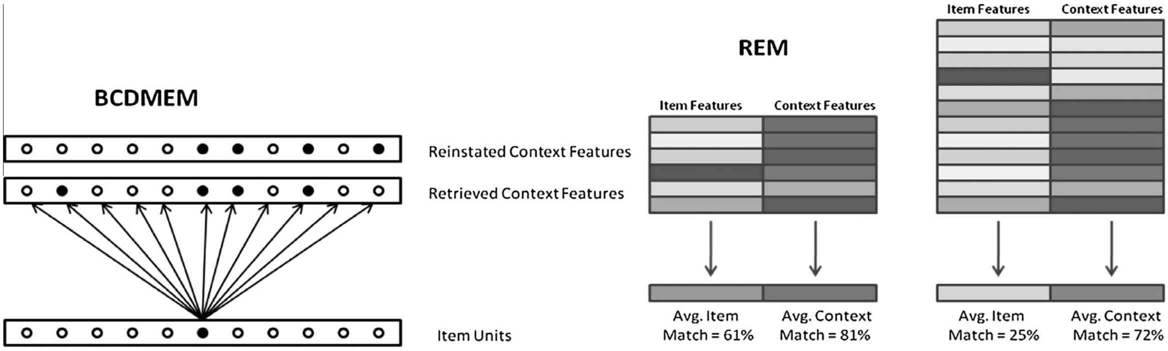


Fig. 1. Visual representation of BCDMEM and REM.

similarity between the retrieval cue and individual traces, the greater the activation of the traces will be. The illustration in Fig. 1 shows a target test trial, and activation of the traces is represented by how dark the traces are, with darkness being positively related to the traces' activations. Hence, there is one trace in memory for which there is strong similarity between the item features in the retrieval cue and item features in the trace. The remaining traces match the retrieval cue by chance and therefore less strongly. The same process occurs for context features in the retrieval cue used to probe memory and context features stored in memory traces. The context matches are about the same on average for the longer and shorter list if context does not change wildly from study trial to study trial.

The familiarity of the target stimulus is the basis of the recognition decision, and it is the average activation of the traces in memory. Since there are a greater number of traces with item features that do not match the item features in the retrieval cue for longer lists, the average familiarity of the targets is lower for long lists than for short lists. Hence, the hit rates are expected to decrease as the number of items studied increases. This is the basis for the REM predictions of a positive LLE.

More recently, it has become clear that another factor needs to be taken into account when considering the effect of the number of items studied on recognition, the number of items tested. Under standard testing conditions, single-item recognition accuracy decreases with increases in the number of items tested (Annis, Malmberg, Criss, & Shiffrin, 2013; Criss, Malmberg et al., 2011; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Malmberg, Lehman, Annis, Criss, & Shiffrin, 2014). This is referred to as output interference and it indicates that modeling the storage of episodic traces during test is necessary, and the result is that any manipulation of list-length is compromised by testing memory.

Fig. 2 shows the REM predictions from Shiffrin and Steyvers (1997) and Criss, Malmberg et al. (2011). Two sets of predictions are shown: one for the no-filler condition and one for the filler condition of Dennis et al.'s experiment. The only difference between the two conditions is that several additional memory traces were stored during the performance of the 8-min filler activity per Turner, Dennis, and Van Zandt (2013, pg. 22 of supplement describing the model simulations). The parameters used to generate these predictions were the same standard set of parameters used in numerous earlier studies. What the predictions show is that REM predicts a small LLE that rapidly decreases in magnitude with increases in the number of items studied in the short-list condition, and that the magnitude of the LLE is virtually nil in the filler condition due an increase in the traces stored during the retention interval.[1]

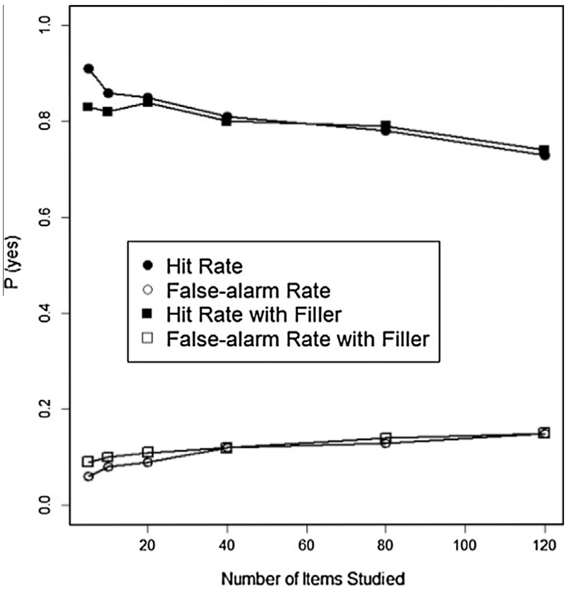


Fig. 2. REM predictions for the filler and no-filler experimental conditions in experiments like those conducted by Dennis and colleagues. Note the dependent variable in Dennis et al.'s (2008) analyses was *d′*. REM predicts about a .19 change in *d′* in filler condition, given the list-lengths of their experiment. This is an exceedingly small change in accuracy representing only a 2% change in the area under the ROC.

Dennis, Lee, & Kinnell's 2008 Experiment

The design of Dennis et al.'s experiment, illustrated in Fig. 3, is a 2 (list-length: 20 vs. 80 words) × 2 (filler task: present vs. absent) × 2 (word frequency: common vs. rare) factorial, with all factors manipulated within-subjects. This basic design has been used in several studies by Dennis and his colleagues, and it is considered by proponents of the context-noise models as the gold standard (Dennis & Humphreys, 2001; Kinnell & Dennis, 2011; Turner et al., 2013., etc.). The "filler task" was an 8-min sliding puzzle activity that was appended to the end of the study lists. The "puzzle activity" was performed in short-list conditions to equate the retention intervals. The "puzzle

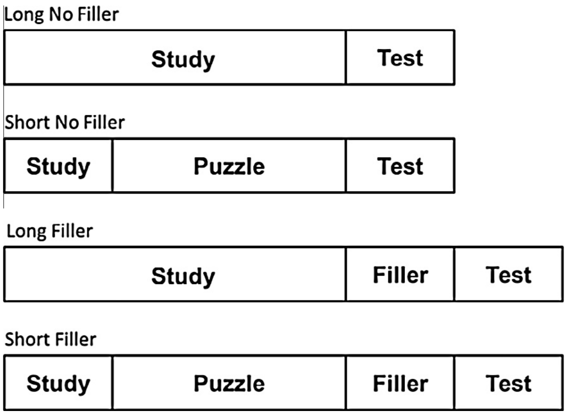


Fig. 3. Design of the experiment conducted by Dennis et al. (2008).

[1] Note that in the simulation, no distinction was made between the traces stored during study and those stored during recognition testing. Hence, even though magnitude of the predicted LLE is quite small, the predictions shown in Fig. 2 are a close approximation, even if they may overstate the predicted effect of list-length on recognition accuracy.

activity" and the "filler task" are the same task (sliding puzzle) with different labels; the terms are used to differentiate the purpose with respect to the conditions. In the filler condition, the intended purpose was to equate the difficulty in reinstating the learning context when memory was tested, whereas in the puzzle condition, the purpose was to equate study-test lag.

Because early tests of the models relied on null hypothesis significance tests that had no value in measuring support for a null LLE, Dennis et al. (2008) conducted their new experiment and advocated a Bayesian data analysis. But rather than directly testing the abilities of BCDMEM and REM to account for their findings and conducting a simple straightforward Bayesian analysis of the data from their experiment, Dennis et al. subjected their data to a newly developed analysis that used a measurement model based on signal detection theory. The use of a Bayesian hierarchical measurement model was motivated by several advantages of using Bayesian statistical analyses over more conventional frequentist analyses. For instance, Bayesian analyses are highly desirable in cases just like the one faced by Dennis et al. because they allow one to make inferences about the plausibility of a null effect, and in this case, they were especially concerned about conclusions about the LLE because some theories may predict a LLE while others may not. (Dennis et al. also presented an incomplete frequentist analysis of their data, but they focused their attention primarily on a Bayesian signal detection analysis of the data from their LLE experiment.) Based on these analyses, they reported "evidence for an absence of a list-length effect". A critical problem is that quantifying support for the null hypothesis within a Bayesian framework requires an analysis of Bayes' factor (Rouder & Lu, 2005), but Dennis et al. did not report Bayes' factor. Therefore, it is unclear how much "evidence" formed the basis of Dennis et al.'s conclusion.

We will turn our attention to Dennis et al.'s Bayesian data analysis below. However, it is first worth briefly considering a set of null hypothesis significance analyses in order to highlight the ambiguous nature of Dennis et al.'s (2008) findings. The dependent variable in Dennis et al.'s analysis was d', and Fig. 4 shows that d' was slightly greater on average in the short-list than the long-list condition regardless of the filler activity conditions.[2] Dennis et al. reported the following:

> In the filler comparison, where contextual reinstatement was encouraged after both the short and long lists, a repeated measures ANOVA yielded a non-significant 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; η_p^2 = .09), suggesting that list length did have an effect on performance. (p. 365)

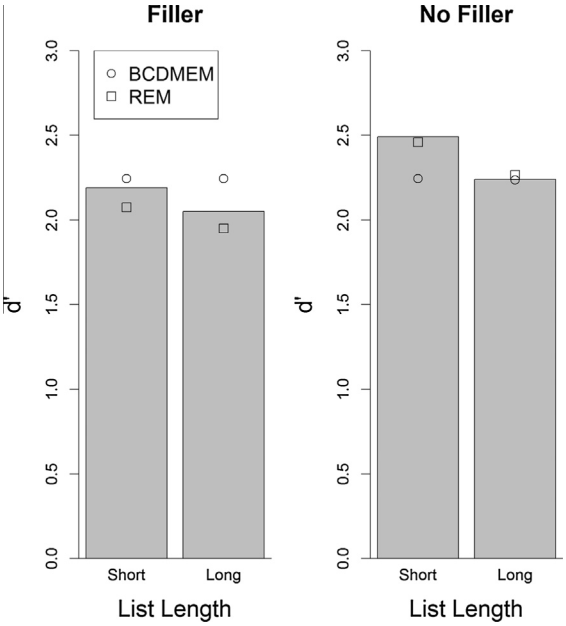


Fig. 4. Results from Dennis et al. (2008). BCDMEM predictions are shown as circles, and REM predictions are shown as squares. See the text for the procedure used to obtain the best fit predictions.

Hence, by the common standard of the scientific community, there was not a significant LLE in the filler condition, but there was a significant LLE in the no-filler condition. A fair interpretation of these results leads one to suspect that there might be an interaction between list-length and the 8-min filler activity. Fortunately, Dennis et al. (2008) published their data, and therefore, we were able to run the analysis.[3] A (list-length: 20 vs. 80 words) × 2 (filler task: present vs. absent) repeated measures ANOVA revealed a main effect of the Filler Task, F(1, 184) = 8.08, p = .005. Subjects had greater recognition accuracy (measured in d') in the no-filler condition (M = 2.36, SD = 0.81) than in the filler condition (M = 2.12, SD = 0.78). It did not detect a significant main effect of list-length (F < 1) or a filler × list-length interaction (F < 1). Hence, Dennis et al.'s one-way ANOVAs support a different set of statistical inferences than the two-way ANOVA that we ran. Whereas their analyses suggest there was a positive

[2] The use of d' as the dependent variable led Dennis et al. (2008) to further problems in the inferences they drew from their data. We will discuss this problem later in the article.

[3] In our frequentist analyses of Dennis et al.'s data, word frequency did not interact with either list-length or the presence of the filler task, and the list-length predictions of the models do not critically depend on this factor. Therefore, all subsequent analyses focus solely on the effects of list-length and the presence versus the absence of the filler task. A three-way repeated measures analysis of variance revealed a main effect of word frequency, F(1, 47) = 100.98, p < .001, η_p^2 = .682, and a main effect of filler task, F(1, 47) = 9.922, p < .05, η_p^2 = .174. Recognition accuracy was greater for low-frequency words than high-frequency words and for words in the no-filler condition than in the filler condition. There was no main effect of list-length, F(1, 47) = 1.339, p = .253. Additionally, none of the interaction effects reached significance. The word frequency × list-length interaction and the list-length × filler interaction were not reliable, F < 1, nor were the word frequency × filler, F(1, 47) = 2.721, p = .106, or the three-way interaction effects, F(1, 47) = 1.141, p = .291.

LLE in the no-filler condition, but not in the filler condition, we did not detect a statistically significant interaction. We believe that, at the very least, these conflicting analyses cause one to be somewhat skeptical of drawing any strong conclusions about the nature the LLE based on Dennis et al.'s findings. Perhaps more importantly these analyses cannot provide support for the null LLE.

Bayesian data analysis

As noted by Dennis et al. (2008), we can rely on Bayesian data analyses to make inferences about the absence or the presence of the LLE. There are many different Bayesian approaches to statistical inference, and statisticians often debate which approach is the best. This debate is not the subject of the present work; therefore, we will consider several different analyses and base our inferences on all of their outcomes. Fortunately, they lead to the same conclusion.

Bayesian parameter estimation

Some researchers advocate a method of *parameter estimation* for the Bayesian analyses when there are not strong prior beliefs about the effect or the models in question (e.g., Kruschke, 2011a, 2011b). Accordingly, one estimates a posterior probability density distribution over the possible differences in the means between two conditions. The dependent variable in Dennis et al.'s primary analysis was d', and the empirical question is whether the difference in d', referred to as μ, is credibly different from 0. A 95% Highest Density Interval (HDI) contains the most likely 95% of parameter values for the differences in recognition accuracy between the long-list and short-list conditions. If the HDI spans 0 we may conclude that a null LLE is credible. If it does not contain 0, and if the mean of the posterior distribution is positive, then we may conclude that a positive LLE is credible (Kruschke, 2010, 2013). We used Wagenmakers, Lodewyckx, Kuriyal, and Grasman's (2010) Bayesian *t*-test model to obtain the posterior.

Fig. 5 shows the results of the Bayesian *t*-test for the filler condition. The empirical question is whether the difference in *d*', referred to as μ, is credibly different from 0. The thick, solid horizontal line at the bottom of the left panel shows that the 95% HDI extends from about .01 to .49, indicating that a null effect of list-length is not among the 95% most credible differences in d'. In other words, there is a credible LLE ranging from almost non-existent to moderately large in the no-filler condition. The left panel of Fig. 6 shows the corresponding results for the filler condition, where the HDI was estimated to be between −.085 and .36. Since 0 falls within the HDI, a null LLE is credible. Hence, the parameter estimation analyses are qualitatively similar to the frequentist analyses reported by Dennis et al. (2008), namely that there is a credible LLE in the no-filler condition and a credible null LLE in the filler condition, these conclusions conflict with those based on Dennis et al.'s signal detection model, which we will discuss shortly.

Hypothesis testing

Another approach to statistical inference is Bayesian *hypothesis testing* where the relative support for multiple hypotheses is evaluated, and our analyses using Wagenmakers et al.'s Bayesian *t*-test also provide the relative support for the null LLE in the two conditions of Dennis et al.'s experiment. The null and alternative hypotheses can be written in terms of the effect size that they predict, H_0: $\delta = 0$, and H_1: $\delta > 0$, respectively. A common approach to Bayesian *hypothesis testing* is to calculate Bayes Factor, BF_{01}, which measures the relative support for the hypotheses we seek to test.[4] If $BF_{01} < 1.0$, the data reduce the relative credibility of the null hypothesis, and if $BF_{01} > 1.0$, then the data increase the relative credibility of the null hypothesis. Often times it is difficult or impossible to determine the Bayes factor because the calculation involves an intractable integral, but if the models are nested, which is the case here, the Savage-Dickey Ratio Test (Dickey, 1971) is a procedure that simplifies the calculation of the Bayes factor. At an intuitive level of understanding, the Savage-Dickey Ratio Test tells us how the data, *D,* alters our prior belief that the effect size equals 0. To obtain the Bayes factor with this technique, the ratio of the height of the posterior and prior probabilities of the effect size at 0 ($\delta = 0$) is calculated (Wagenmakers et al., 2010).

The prior probability density of the effect size, δ, is depicted as the dashed line in the right panel of Figs. 5 and 6 for the no-filler and filler conditions, respectively. Note, our choice of priors respects a wide range of prior beliefs about the size of the effect that list-length has on item recognition by assuming that prior to the experiment a null LLE is more likely than any other magnitude of LLE, but also allowing for the possibility of a positive LLE. The solid line plots the posterior distribution over δ indicating a range of credible beliefs after the data have been collected. For the no-filler condition in Fig. 5, $BF_{01} = .43$ which indicates that the data are approximately 2.33 times more likely given a positive LLE than given a null LLE, whereas the data were found to be 1.6 times more likely under the null hypothesis than under the alternative hypothesis for the filler condition in Fig. 6.[5]

Summary of findings

Although Dennis et al. declared that they found "evidence against a list length effect", they did not report an

[4] We are ignoring the impact of the priors placed on H_0 and H_1. In essence, we are assuming that they are equally likely prior to receiving Dennis et al.'s data. The layperson for instance may place a greater prior on H_1 because he believes that long lists are more difficult to remember than short lists. Hence, assuming equal priors makes it more difficult to find support for the H_1 than the layperson might desire.

[5] Another way to compute Bayes Factor is to compute it directly from the result of the paired-sampled *t*-test Rouder, Speckman, Sun, Morey, and Iverson (2009). When this computation is performed a $BF_{01} = 2.97$ is found for the filler conditions and a $BF_{01} = .84$ is found for the no filler condition. This analysis indicates that the data are about 3 times more likely under the null hypothesis than under the alternative hypothesis in the filler condition, and that the data are approximately 1.4 times more likely under the alternative hypothesis than under the null hypothesis in the no-filler condition.

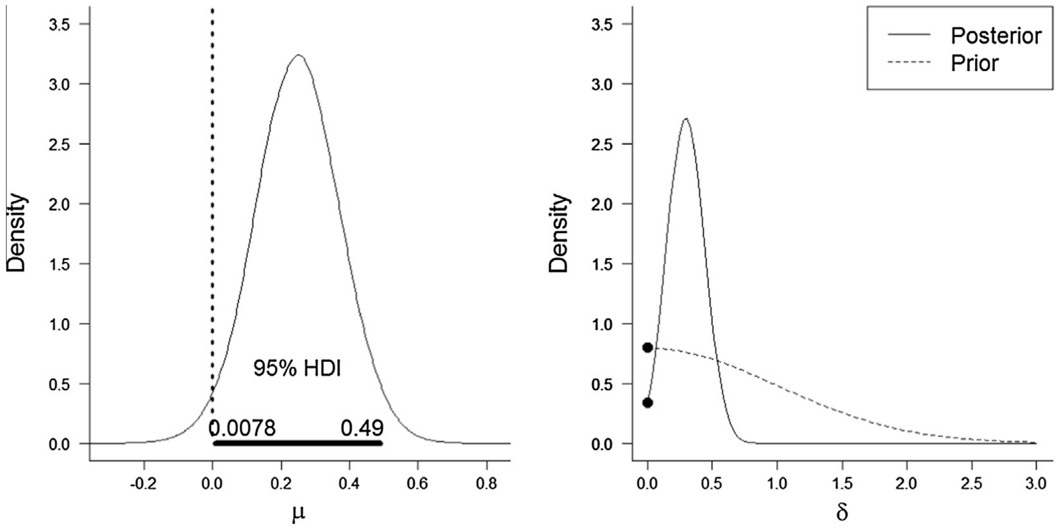


Fig. 5. Analysis of the no-filler condition of Dennis et al. (2008). The left panel plots the posterior density distribution of the mean difference between list-length conditions. Zero falls slightly outside the 95% HDI. The right panel plots the prior and posterior density distributions over the effect size. The density over effect size equal to 0 is reduced slightly after the data are taken into account.

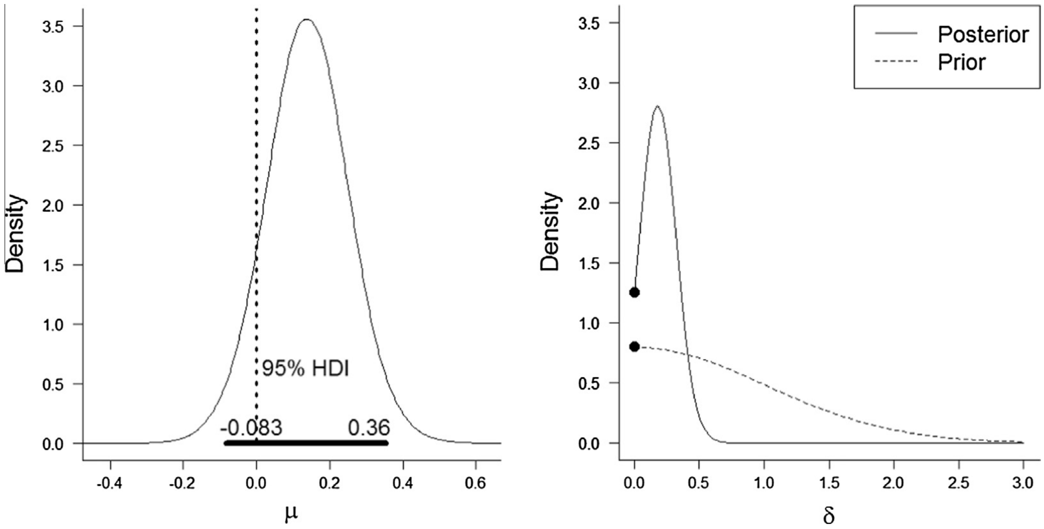


Fig. 6. Analysis of the filler condition of Dennis et al. (2008). The left panel plots the posterior density distribution of the mean difference between list-length conditions. Zero falls within the 95% HDI. The right panel plots the prior and posterior density distributions over the effect size. The density over effect size equal to 0 is increased slightly after the data are taken into account.

analysis that weighed the relative support for and against the LLE. The data provide support for a positive LLE in the no-filler condition and a null LLE in filler condition, but the evidence in both cases is equivocal; Dennis et al.'s findings have little diagnostic value one way or the other. For example, according to Jeffreys (1961), a Bayes Factor between 1 and 3 is "worth no more than a bare mention", while Wagenmakers et al. (2010) interpret it as "anecdotal evidence." Therefore, although the Bayesian analysis of the data are consistent with Dennis et al.'s null-hypothesis significance tests, their results have a considerable amount of variability relative to any LLE that may be present, and therefore they have only a little diagnostic value according to published guidelines.

Dennis, Lee, and Kinnell's Bayesian hierarchical signal detection analysis

In contrast to these conclusions, Dennis et al. (2008) expressed a strong preference for models that predict a null effect derived from their Bayesian mixture model analysis:

> 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, challenges item noise accounts and suggests that "recall-like" processes play no substantive role in yes/no recognition.
>
> [(p. 373)]

Note that their conclusions were not based on the analyses that we have discussed so far, which indicate that there is no firm basis for any strong conclusions, especially about the null LLE. Rather, Dennis et al.'s conclusions were based on the outcome of a Bayesian mixture model analysis of individual differences. The apparent disagreement between what our analysis tells us, and what the Dennis et al. Bayesian mixture-model analysis tells us is bothersome and important to investigate.

Whereas the analyses that we reported focused on parameter estimation and hypothesis testing, Dennis et al. focused on model comparison methods conducted within a theoretical framework based on a Bayesian hierarchical model of signal detection theory (SDT; Green & Swets, 1966). The formal details of their analyses are presented in Appendix A. First, Dennis et al. described a model of the null effect, which they referred to as the error-only model, and a model of a signal-plus-noise, which they referred to as the effect-plus-error model. Their goal was to determine whether the observations from their experiment were characteristic of an error-only model (null effect) or an effect-plus-error model (positive LLE). To compare the error-only model to the effect-plus-error model, a set of parameters values was sampled randomly from the error-only model and a sample of parameters values was sampled randomly from the effect-plus-error model, the posterior probabilities of the models were computed, and the winning model was chosen. This process was carried out many thousands of times in order to obtain sets of parameters spanning the parameter spaces of both models, and this produced the posterior probability distribution over the rate at which the error-only model was selected over the effect-plus-error model. Based on their analysis, Dennis et al. concluded that the null effect model was strongly preferred over the signal-plus-noise model.

Dennis et al.'s strong conclusions are at odds with every other analysis of their data, including the "model free" NHST analyses conducted by Dennis et al. (2008) and those we reported in the earlier sections of the article. This led us to scrutinize the Bayesian hierarchical mixture model. A first step was to conduct a standard posterior predictive check of the models Dennis et al. analyzed. The posterior obtained from the analyses not only provides the information about the rate at which models were preferred, but also about information about the predictions of the models they tested. The goal of the posterior predictive analysis is to determine whether the models on which the Bayesian hierarchical analysis is based correspond to the data they attempt to describe. If the models do not provide accurate accounts of the data they attempt to describe, the relative abilities of the models to describe the data is not informative.

For the posterior predictive analysis, we simply obtained the differences in d' values from the posterior of each model and compared them to a resampled set of data from the filler and no-filler conditions of Dennis et al.'s experiment. The d' difference scores represent the change in discriminability resulting from an increase in the number of items studied. The resampled set of data was obtained by sampling the d' difference scores from normal distributions with means and standard deviations corresponding to Dennis et al.'s observations many times and fitting a smoothing function over them. Hence, resampled data are the d' difference scores that we would expect to observe if Dennis et al.'s experiment was conducted many times or with many subjects.

The result of the posterior predictive check is shown in Fig. 7. The models analyzed by Dennis et al. fail miserably. First, neither the error-only nor the effect-plus-error model predicts the LLE in the no-filler condition. The predicted means of the posterior distributions are less than the means of the data by a large margin. Not surprisingly, the effect-plus-error model is slightly more successful, but there is still a significant amount of the posterior that does not overlap with the data. Second, the variability in the effect-plus-error model is much greater than the variability in the data, and the error-only model fairs only a little better. Hence, Dennis et al.'s conclusion that the error-only model provides a much better account of their observations than the effect-plus-error model was based on a comparison of two models that do not predict their findings.

The underlying logic of Dennis et al's. analyses was that the error-only model is roughly analogous to a context-noise model and the effect-plus-error model is roughly analogous to an item-noise model. In fact, Dennis et al. drew conclusions from their analyses that went beyond the models that they tested, when they stated that their findings were consistent with context-noise models and disconfirmed item-noise models. However, they did not test models like REM and BCDMEM, and therefore, their analyses are only anecdotally related to the issue of whether their data are more characteristic of item-noise or context-noise models. Furthermore, it is not clear how useful their approach is for discriminating between such models. In order for Dennis et al.'s hierarchical analysis to have anything meaningful to say about the viability of item-noise and context-noise models, their hierarchical model must be able to discriminate between data generated by item-noise models and context-noise models with some high degree of accuracy. If it cannot, then Dennis et al.'s conclusion that their analyses support BCDMEM and disconfirms REM is unjustified.

To determine whether Dennis et al.'s hierarchical analysis can discriminate between data from BCDMEM from data from REM we conducted two simulations in order to generate simulated subjects from REM and BCDMEM. Each set of data was meant to simulate one subject in Dennis et al.'s experiment. Since REM and BCDMEM are stochastic models, each sample of 48 subjects is different. Here we present three representative simulated data sets in order to illustrate the robustness of our analyses and our conclusions. We were careful to judiciously choose parameters for each model from which to obtain the simulated data sets. For REM, we mostly used the parameters values taken from the literature to simulate 48 datasets for

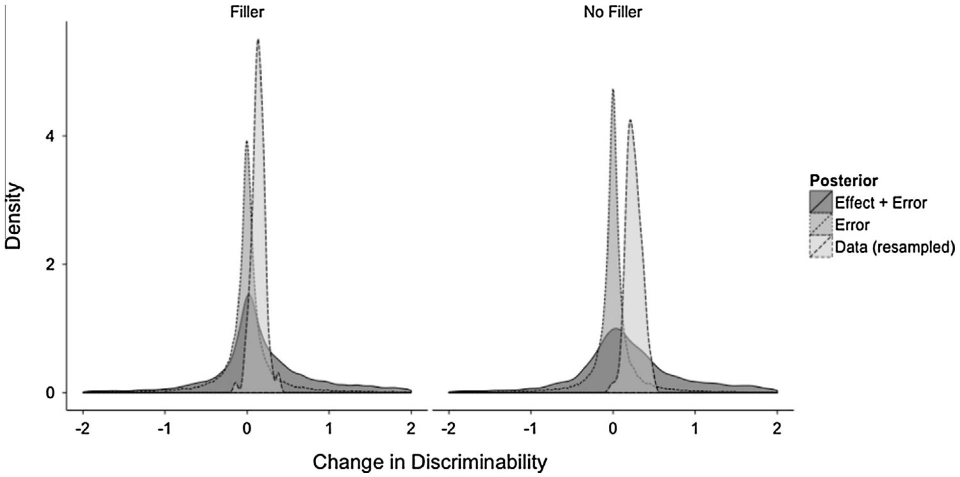


Fig. 7. Posterior predictive check for the error-only and effect-plus-error model.

Dennis et al.'s experiment. This makes sense because these parameters are known to account for findings from a wide variety of experiments, and there is no reason that we could justify for deviating from prior research. The following REM parameters were fixed across all conditions: $g = .4$, $c = .7$, $u^* = .04$, $t = 15$, and $w = 20$.

To achieve a good fit of REM to the mean accuracy (d') of Dennis et al.'s data, we adjusted how well encoded the memory traces were by manipulating REM's t parameter. This is a standard way of modeling strength of encoding in REM (Shiffrin & Steyvers, 1997; Xu & Malmberg, 2007). Specifically, we varied t from 10 to 18 in increments of one, recording the log-likelihood on each run. The absolute goodness of fit obtained was measured with the maximum log-likelihood statistic (LL = −235.61). We found the best fitting value to be $t = 15$. Hence, the fit is obtained by using parameters values for REM obtained from the literature and scaling accuracy in order to achieve the proper level of overall accuracy. In order to model the difference in retention intervals, we assumed that five traces per minute were stored during the 8-min. filler activity (as in Turner et al., 2013). According to both this analysis and visual inspection of Fig. 4, the fit we obtained was closely in line with data.

We obtained the same number of datasets of d' difference scores corresponding to BCDMEM. Standard BCDMEM parameter values are $s = .02$ and $v = 200$. For other parameters of BCDMEM, there is less consistency in the literature. We found $r = .8$ and $p = .1$ to provide reasonable fits. It should be noted that there are many values of r and p that we considered that resulted in the same fit. Thus, our choice of r and p was not an influential factor in the outcome of the fits. We also maintained Turner et al.'s (2013) assumption that the parameter d varies between the long-list no-filler condition and the other three conditions. Accordingly, we first obtained the value of d that minimized the RMSD by using the quasi-Newton optimization method in R for the long-list no-filler condition and then obtained a single value for d for all other conditions using the same method. We found the best fitting value for the long-list no-filler condition and the other conditions to be identical, $d = .54$ (LL = −237. 91). According to both this analysis and visual inspection of Fig. 4, the fit we obtained was closely in line with data.[6]

With these models, we simulated forty-eight subjects from each model with the best fitting parameter values given above. Because both REM and BCDMEM are probabilistic models, each simulated set of 48 subjects will differ from another. In order to demonstrate the variability between simulations, we generated three independent sets of 48 simulated subjects each keeping parameter settings constant across sets. We then used these simulated data to test whether Denis et al.'s analyses can discriminate between data sets known to be generated by item-noise and context-noise models.

These simulated data were *not* used to inform the priors of Dennis et al.'s model. We used the same reasonable uniformed priors that Dennis et al. used. We simply asked, could their signal detection hierarchical model actually discriminate between data sets generated by an error-only model versus an error-plus-effect model? In other words, can conclusions drawn from the SDT hierarchical model be used to draw conclusions about REM and BCDMEM, as Dennis et al. suggest?

To answer this question, the simulated subjects were used as the data in Dennis et al.'s mixture model of individual differences. One-hundred and fifty-five thousand MCMC samples were generated using the JAGS software. The first 5000 samples were discarded, and after checking for convergence, the chains were collapsed. We repeated

[6] AIC and BIC statistics were also computed and found a similar pattern of relative goodness of fit. The reason that a single d value is obtained when BCDMEM is optimized is because BCDMEM is required to, as REM was, fit the data from both filler conditions, and the mean for the long-list no-filler condition is nearly equal to the pooled mean of the other conditions. Hence, when scaling the accuracy of the long-list no-filler condition to the accuracy of the other conditions the parameter search produced a compromise for the best fit; there is no effect of list length or the retention interval.

this process for each simulated set of data. The important posterior is over θ, the rate at which individuals are assigned to the effect-plus-error model. When $\theta = 0$, the probability of assigning an individual to the effect-plus-error model is 0, while the rate of assigning an individual to the error-only condition is 1.

If Dennis et al.'s model can discriminate between data generated by REM and BCDMEM, then the model should assign data generated by REM to the effect-plus-error model and data generated by BCDMEM to the error-only model. Each row of Fig. 8 shows the density over θ for those data generated by either REM or BCDMEM for each simulated set of 48 subjects. There is variability within each set for both BCDMEM and REM. Panel A shows the results for the no-filler condition for the first set of simulated data. The surprising result was that most of the density is concentrated around low values of θ for the subjects obtained from REM simulations indicating a high rate of assignment to the error-only model. This is because REM naturally predicts a small LLE, and this is actually more difficult for the effect-plus-error model to predict. When simulated subjects generated by BCDMEM are used, the posterior density is peaked when θ is approximately equal to .60. The individual differences model is about as likely to assign data from BCDMEM to the error-only model as to the effect-plus-error model. Panels C and E show the density over θ for the next two simulated data sets. In these cases, the posterior is almost evenly distributed over θ, and Dennis et al.'s model is unable to classify REM-simulated subjects with any degree of certainty. The mixture model analyses revealed a similar result for the data generated by BCDMEM, although the posterior is more tightly centered on $\theta \sim .60$, indicating that the model is quite uncertain whether the data were generated by an error-only model. In summary, Dennis et al.'s model is not able to discriminate between data sets coming from error-only and error-plus-signal models, and sometimes, simulated data from REM is more likely to be assigned to a model predicting a null LLE than simulated data from BCDMEM.

Panel B of Fig. 8 shows the mixture model analysis of the simulated subjects generated by BCDMEM and REM for the filler condition. In this condition, our data analyses indicated more support for a null LLE than for a positive LLE. Although the data are different, the same problems for Dennis et al.'s mixture model that we found in the no-filler condition are also found here, but they are compounded because BCDMEM and REM make similar predictions for the filler condition. In panel B, an almost non-existent LLE was observed in the simulated data produced by REM. However, Dennis et al.'s mixture model mistakenly categorized the data as coming from an error-only model. For the simulated data produced by BCDMEM, the model did not return conclusive evidence

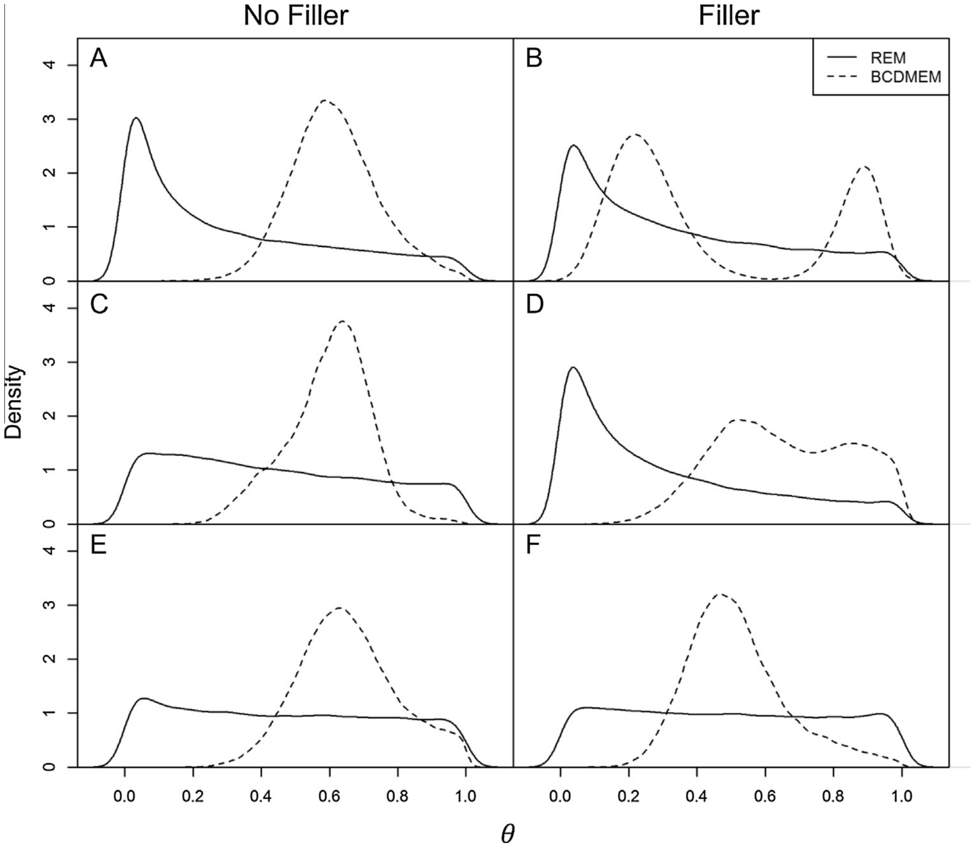


Fig. 8. Shows the density as a function of the rate parameter θ. When $\theta = 0$, the probability of assigning an individual to the error-only model is 0, while the rate of assigning an individual to the error-plus effect model is 1. The different lines refer to the model used to generate the data.

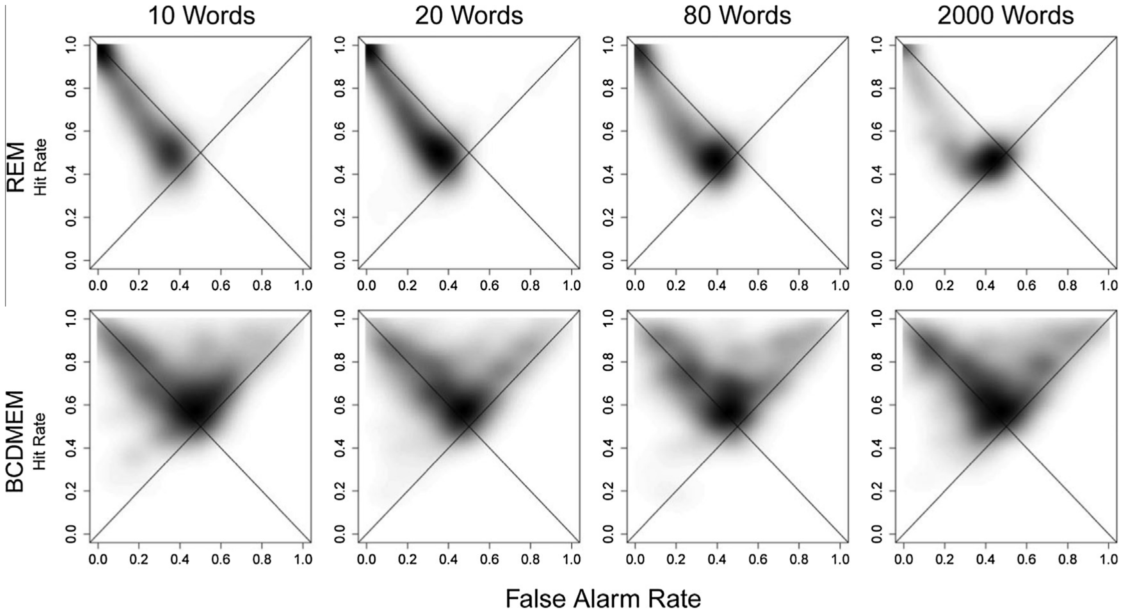


Fig. 9. Figure from Turner, Dennis, & Van Zandt (2013) showing that BCDMEM predicts that there should considerable variability among subjects, whereas REM more consistently predicts small LLEs. The prior predictive density under four different list lengths for REM (top panels) and BCDMEM (bottom panels): 10 (left column), 20 (middle left column), 80 (middle right column), and 2000 (right column) items. Darker regions indicate higher density. BCDMEM = bind cue decide model of episodic memory; REM = retrieving effectively from memory model.

for either an error-plus effect model or an error-only model, as indicated by the bimodality of the posterior distribution. This is again due to the high degree of variability in BCDMEM predictions that we identified earlier. Panel D shows the results on the next set of simulated data. The mixture model assigned BCDMEM to the error-plus-effect model while assigning REM to the error-only model. Panel F shows the density over θ for REM to be widely distributed while the density for BCDMEM was centered over $\sim$ .50, indicating that mixture model is equally likely to assign data from a context-noise model to the error-only as to the error-plus-effect model.

These analyses show that Dennis et al.'s mixture model cannot reliably discriminate between item-noise and context noise models of memory. It is somewhat surprising that Dennis et al.'s SDT hierarchical model often misclassifies BCDMEM data by assigning it to the effect-plus-error model. This is because there is a considerable amount of variability in BCDMEM predictions, and when it predicts a LLE by chance, it is more characteristic of the effect-plus-error model than the error-only model. Fig. 9 was reported by Turner et al. (2013). It shows the variability of predictions of REM and BCDMEM. Their plots clearly show that BCDMEM predicts a wider range of LLEs than REM does, many of which are positive LLE, which Dennis et al.'s model misclassifies as coming from an error-plus-effect model.[7]

This analysis of Dennis et al.'s (2008) Bayesian hierarchical mixture model indicates that list-length experiments, like the one purported by Dennis et al. (2008) do not provide critical empirical tests of REM and BCDMEM. For this reason, Dennis et al.'s conclusion that the results of their experiment support BCDMEM and disconfirm REM is erroneous.

General discussion

Dennis et al. claimed they found evidence against a LLE for recognition memory. Our analyses of their observations indicate that there may be a small LLE that diminishes with increases in the retention interval, but we did not find strong support for either the null or alternative hypothesis. Hence, Dennis et al.'s conclusions about the presence or absence of the LLE are at least questionable and quite possibly incorrect. Moreover, the ambiguous nature of Dennis et al.'s observations calls into question the utility for testing models of recognition memory using the experimental designs advocated by Dennis and colleagues. The problem is that it is very difficult to distinguish between a model that predicts no effect and a model that predicts a small effect given data that are somewhat noisy. This problem is compounded by the fact that their design uses a different retention interval for each condition of the experiment making the results uninterpretable (see Fig. 3). Other experimental designs allow for more precise measurements, such as manipulations of category length, and the results obtained from them have shown clear positive LLEs, especially for non-word stimuli (Criss & Shiffrin, 2004a, 2004b). In addition, we discovered that Dennis et al.'s Bayesian hierarchical mixture model of individual differences is insufficient for discriminating between data simulated by BCDMEM and REM. We found no strong evidence that support the conclusion that their observations

[7] It should be noted that Turner et al. (2013) did not model the effect on the retention interval of the "puzzle activity", and therefore these predictions over-state REM's predicted LLE.

are consistent with context-noise models of recognition memory and inconsistent with item-noise models.

The list-length effect: Is it theoretically important?

The effect of the number of items studied on recognition memory has been extensively investigated for over a decade, and for over a decade some have claimed that these findings are easy for BCDMEM to predict and difficult for REM to predict (Dennis & Humphreys, 2001). If so, then perhaps the LLE is theoretically important. However, our conclusions are different, and they arise from a systematic investigation. Here we provided a comprehensive statistical analysis of the data driving the debate. Several analyses were reported in order to get the clearest possible picture of the empirical result in question. Because our statistical analyses did not support Dennis et al.'s conclusion that their findings are easy for context-noise models and difficult for item-noise models to explain, we next investigated the models they used to form their conclusion. This investigation identified critical limitations of the Bayesian mixture model to measure what it purported to measure; namely Dennis et al.'s conclusions were based on a model that cannot accurately discriminate between data generated by context-noise models and item-noise models.

Our analyses support two conclusions: First, Dennis et al.'s (2008) data are not sufficient for competitively testing models of recognition memory because the effects are small and the predictions of the models are somewhat similar. Second, BCDMEM cannot account for Dennis et al.'s findings showing a positive LLE at short retention intervals without including a parameter value associated with the "reinstatement of context." For example, Turner et al. (2013) modeled these data by assuming one context reinstatement parameter for the long-list no-filler condition and another for all other conditions. From that perspective, the retention interval presents a confound rather than a control. In addition, we found that BCDMEM has a difficult time simultaneously predicting the effects of list length and retention interval. For instance, d' was lower after the long retention interval in the filler condition, but Fig. 4 shows that BCDMEM predicted no effect of the filler task. It is important to note that we did not pick the model that BCDMEM claims to provide a superior account of these data; the account was actually proposed by Turner et al. (2013). We simply implemented their model.

We also found that when their experiment is modeled in a manner consistent with what is known about recognition memory, the results are not problematic for traditional models of memory, like REM, that predict interference from representations of the items studied and tested. Of course, we did not conduct a similar set of analyses on every set of data reported by Dennis and his colleagues over the last decade, but the data from Dennis et al. (2008) are sufficiently representative of the experimental design in question to make useful generalizations. One specific example of data that is problematic for context-noise models is the Kinnell and Dennis (2011) data showing a positive LLE for fractals and faces but not landscape pictures. These findings may be deemed important with replication but there are no inherently obvious ways to model these stimuli (which are different from words) with context-noise models. Turner et al. (2013) modeled these data (see Study 2 in their paper) and concluded that BCDMEM fits better than REM, but given that 2 of 3 stimulus conditions show a positive LLE and BCDMEM cannot predict a LLE without additional assumptions[8], we must assume that this is a case where BCDMEM may fit better in quantitative terms but fails to fit the qualitative pattern of the data. These so called 'better' fits are due to three things (1) high variability in BCDMEM as we covered already (2) the ancillary assumption that the puzzle-filler task adds traces but the puzzle task does not (this is critical for the Dennis et al. (2008) data), and (3) because they do not fit the qualitative pattern of the data.

For the data in question here, we did not need to address the issue of how contextual representations affect recognition memory within the REM framework. We have, however, considered how contextual dynamics affect recognition elsewhere (Criss & Shiffrin, 2004a, 2004b, 2005; Criss, Malmberg et al., 2011; Lehman & Malmberg, 2009; Malmberg & Shiffrin, 2005; Murnane, Phelps, & Malmberg, 1999), but for present purposes we see no compelling reason why REM's predictions would change in a significant way if the additional complexity were added. And although we considered different ways that BCDMEM could be modified to better handle the data from Dennis et al. (2008), we did not go to great pains to implement these models either. The reason is because the obvious modification of BCDMEM would be inconsistent with the model described by Turner et al. (2013), who assumed that the ability to reinstate the learning context is unaffected by the length of the retention interval. In this sense, the ability to reinstate context, according to BCDMEM, is not conceptualized as a general source of forgetting. Rather, it suggests that once a temporary contextual representation is lost, the reinstatement of it is subject to error, but the amount of error does not increase over time. Such a model is, for instance, consistent with a multiple store buffer model that assumes once the learning context is removed from a short-term store that the damage is done (Atkinson & Shiffrin, 1968; Lehman & Malmberg, 2013).

One unexpected finding in our analyses was that at times Dennis et al.'s hierarchical model mistaken classified data generated from the context-noise version of the error-only model as having been generated by the REM version of the error-plus-signal model. To see why note that the BCDMEM does not predict a LLE in principle, but in reality the model has sufficient noise that it is quite capable of predicting a small positive (or negative) LLE. BCDMEM while seemingly very constrained in terms of the verbal theory (e.g., predicts no LLE) is actually minimally constrained in terms of the mathematics (e.g., it predicts a wide distribution of possible values of HR and FAR, centered on a null LLE). In contrast, REM is quite constrained in both theory and mathematics in that it predicts a more narrow distribution around a small positive LLE.

[8] The supplementary material to Turner et al. (2013) does not specify what assumptions lead BCDMEM to predict a positive LLE. Neither can we think of a way for BCDMEM to provide a better fit than REM when positive LLEs are observed.

The range of predictions were reported by Turner et al. (2013) and we reprint their figure in Fig. 9, which shows that the predictive space of BCDMEM is hugely variable (the entire V shaped blob) whereas REM is fairly constrained. Hence, when the error-only model generates data in accord with BCDMEM it produces differences in d' with a great deal of variability, and when it by chance produces a positive LLE, the Dennis et al.'s model is likely to assign it the error-plus-signal model.

Conclusion

Our analyses highlight the interaction between statistical inference and cognitive modeling and our understanding of data and theory. They indicate that the LLE effect for single-item recognition is likely to be quite small and approach nil with increases in retention intervals. Our analysis of Dennis et al.'s data indicates that they are unsuitable for testing item-noise models like REM and context-noise models like BCDMEM. Even though those data do not strongly favor one model over the other, there are other findings in the literature that do. How likely are the models given those data? That is the $64,000 question, and Bayes factor does not provide the answer. To answer this question, we need prior probabilities on the models themselves, which may seem difficult to obtain, and indeed assigning specific probabilities to these models with any degree of confidence is foolhardy. But here is what we do know, and what we know is true may be used to assign relative probabilities to the models.

BCDMEM is disconfirmed by several findings including the effects of list composition on the word frequency effect (Dorfman & Glanzer, 1988; Malmberg & Murnane, 2002), feature frequency effects (Criss & Malmberg, 2008; Malmberg, Steyvers, Stephens, & Shiffrin, 2002), word frequency effects (Hemmer & Criss, 2013), output interference (Annis et al., 2013; Criss, Malmberg et al., 2011; Koop, Criss, & Malmberg, 2015; Malmberg et al., 2012), the interaction between item similarity, word frequency and repetitions (Criss & Shiffrin, 2004a, 2004b; Malmberg, Holden, & Shiffrin, 2004), and item interference in multiple list recognition memory experiments (Criss, 2006; Criss & Shiffrin, 2004a, 2004b). No solutions to these challenges have been forthcoming; thus, we think it is unlikely that current versions of BCDMEM can easily be modified to accommodate them. Moreover, BCDMEM only accounts for item recognition and does not account for other memory tasks that are critical to understanding memory, such as free and cued recall (e.g., Aue, Criss, & Fischetti, 2012; Criss, Aue, & Smith, 2011; Lehman & Malmberg, 2013). Hence, even with minimal empirical constraints, the BCDMEM architecture is strained.

On the other hand, REM provides reasonable accounts for the findings of Dennis et al. (2008) and many of the aforementioned findings that BCDMEM cannot account for. For instance, REM is systematically couched within a theoretical framework that has been formally shown over nearly 50 years to account for scores of factors influencing recall, recognition, lexical decision, perceptual identification, judgments of frequency, associative recognition, and source memory, among others (Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984; Mensink & Raaijmakers, 1988; Schooler, Shiffrin, & Raaijmakers, 2001; Shiffrin & Steyvers, 1997; Shiffrin & Steyvers, 1998; Wagenmakers et al., 2004; Criss & Koop, in press; Criss & Shiffrin, 2005; Lehman & Malmberg, 2009; Lehman & Malmberg, 2013; Malmberg, 2008; Nelson & Shiffrin, 2013). On these observations, we feel confident that the prior probability of the REM model is greater than the prior probability of BCDMEM. If we combine these priors on the models with the analyses we report, we find that our beliefs are unchanged.

Appendix A

SDT models commonly assume the evidence, E, that an item was studied is a continuous, random variable. On each trial, for instance, E may be drawn from a normal distribution that corresponds to one of the two stimulus classes: the foil distribution consists of noise, and the target distribution consists of a signal plus noise. In recognition memory, the signal represents the prior occurrence of a target and the noise represents background familiarity with the items. The standardized difference between these two distributions is known as d' and is positively related to accuracy. After E is obtained, the subject compares the evidence that the item was studied to a criterion, c. If $E \leqslant c$, then the subject responds "New." If $E > c$, then an "Old" response is made. If the variances of the target and foil distributions are equal, both d' and c can be estimated by calculating the difference between standardized hit and false-alarm rates (Macmillan & Creelman, 1991),

$$d' = z(\mathrm{HIT}) - z(\mathrm{FA}).$$

The criterion, c, can be estimated as the midpoint between standardized hit and false-alarm rates.

$$c = \frac{z(\mathrm{HIT}) + z(\mathrm{FA})}{2}.$$

Hence, the hit and false-alarm rates correspond to the following integrals:

$$P(\mathrm{HIT}) = \int_c^\infty p(E|\mathrm{Old})dE,$$

$$P(\mathrm{FA}) = \int_c^\infty p(E|\mathrm{New})dE,$$

where c, is the criterion, and $p(E|\mathrm{Old})$ and $p(E|\mathrm{New})$ are the probability density functions of the evidence given that the item was a target or a foil, respectively. These integrals are equivalent to the normal cumulative distribution function, Φ, parameterized in terms of d_i and c_i, where i corresponds to the ith subject.

$$h_i = \Phi\left(\frac{d_i}{2} - c_i\right)$$

$$f_i = \Phi\left(-\frac{d_i}{2} - c_i\right)$$

To implement a model in which the Old and New distributions do not have the same variance, the false-alarm rates are defined as the following:

$$f_i = \Phi\left(\left(-\frac{d_i}{2} - c_i\right)\tau^{-1}\right)$$

where τ corresponds to the ratio of the target and foil variances, $\sigma_{\text{New}}/\sigma_{\text{Old}}$. When $\tau < 1$, the variance of the Old distribution is greater than the variance of the New distribution. Usually, single-item recognition ROCs indicate that $\tau \approx .8$ (Green & Swets, 1966).

For each subject, i, both d_i and c_i are normally distributed random variables,

$$d_i \sim Gaussian(0, 2),$$

$$c_i \sim Gaussian(0, 0.5).$$

Since the data consist of hit and false-alarm rate counts, the data are assumed to be binomially distributed,

$$H_i \sim Binomial(T, h_i),$$

$$F_i \sim Binomial(D, f_i),$$

where T, is the number of targets trials, D is the number of foil trials, and h_i and f_i are the hit rates and false-alarm rates for the ith subject, respectively.

Dennis et al.'s error-only model and effect-plus-error model

The signal detection assumptions are fairly straightforward and conventional. This is not the case concerning how two different signal detection models were implemented in order to characterize individual differences in recognition memory performance.

Dennis et al.'s *error-only model* assumes that any changes in performance between short- and long-lists for each subject, $\Delta d_i = d_i^A - d_i^B$, were generated by random error. This corresponds to a null LLE prediction. This is represented as Δe and the assumption is that it is distributed as a 0 mean Gaussian with unknown variance,

$$\Delta e \sim Gaussian(0, \lambda_e).$$

By convention (Spiegelhalter, Thomas, & Best, 1996), λ_e is distributed as an inverse gamma with very low rate and shape parameters,

$$\lambda_e \sim InverseGamma(.001, .001).$$

This distribution approximates the Jeffrey's prior (Jaynes, 1968) and is used because of its "uninformative" nature.

Their *effect-plus-error model* assumes that changes in performance between the short- and long-lists are due to both systematic and random error, and therefore, this model predicts a LLE. The effect component in the graphical model is represented as f^f and is assumed to follow a gamma distribution with shape, α, and rate, β,

$$f^f \sim Gamma(\alpha, \beta).$$

Modeling the effect component as a gamma distribution ensures that the systematic error will always be positive since the gamma distribution is supported on the semi-infinite interval, $[0, \infty)$. Dennis et al. (2008) assumed the following distributions for the hyperpriors, α and β,

$$\alpha \sim \text{Exponential}(1),$$

and

$$\beta \sim \text{Gamma}(0.1, 0.1).$$

The random error component, f^e, is modeled in the same way as in the error-only model – as a 0 mean Gaussian with unknown variance

$$f^e = Gaussian(0, \lambda_e),$$

where again an uninformative prior is assumed for the variance, λ_f,

$$\lambda_f \sim InverseGamma(.001, .001).$$

The random error component allows this model to produce positive, null, and negative LLEs. According to the effect-plus-error model just described, the change in performance between short- and long-lists, $\Delta d_i = d_i^A - d_i^B$, can be modeled as the sum of the systematic and random error components,

$$\Delta f = f^f + f^e.$$

Hence, according to this model each subject, I, has a propensity to be characterized by the error-only model and an inverse propensity to be characterized by the effect-plus-error model. In other words, some subjects produce a positive LLE and others do not.

Model selection

Having described both models under consideration, we now turn to model selection, where the goal is to infer which model is more likely given the data. In the graphical model, a variable, x, is used to select between each model. x follows a Bernoulli distribution with rate, θ,

$$x \sim Bernoulli(\theta).$$

Thus, x can take on either a value of 1 or a value of 0. When $x = 0$, the change in performance between lists, Δd_i, is modeled as random error, Δe. When $x = 1$, changes in performance are assumed to be generated by a combination of systematic and random error, Δf. Thus,

$$\Delta d_i = \begin{cases} \Delta e, & x = 0 \\ \Delta f, & x = 1 \end{cases}$$

Hence, the posterior of interest concerns, θ, or the rate at which the error-only model is preferred over the effect-plus-error model. This is the measure of the propensity for an individual subject to belong to the error-only population versus the effect-plus-error population.

Appendix B

Fig. 1B depicts Wagenmakers et al.'s t-test for Bayesian analysis as a directed acyclic graph. The graph consists of plates, nodes and edges. Each plate consists of the observed data for each subject. In our case, there are 48 subjects, and each subject has a score, X_i, that is the difference in their d' between the long- and short-list conditions. Positive differences indicate that accuracy on the long list is lower than accuracy on short lists. Each node lying outside of the plate represents a parameter, and edges denote the conditional relationships between each. Differences in d', X_i, are conditional on two parameters. There are several different types of nodes. Circular nodes represent continuous variables while square nodes are discrete variables. In this case, all the parameters are continuous random variables. Unshaded and shaded nodes represent unobserved and observed variables, respectively. Here only the differences in d' are observed, and the remainder are the parameters that will be estimated. Variables that are enclosed in concentric circles are defined by a deterministic function of other variables. For the first analysis, our main interest is in, μ, which is the difference in d' between the long and short conditions, and it is computed directly from the values sampled from two distributions: δ, which is the effect size, and σ, which measures how variable the observations are.

In this case, μ is the difference in d' scores that were obtained from the short-list and long-list conditions. We represent the current data with a normal distribution because any difference in two d' scores may take a real value between $-\infty$ and ∞, and because the data reported by Dennis et al. appear by visual inspection to be approximately normally distributed.[9] Hence, X_i represents a difference in d' between the long- and short-list conditions for the ith subject, and it is distributed normally with mean, μ and variance, σ^2,

$$X_i \sim Normal(\mu, \sigma^2). \tag{1}$$

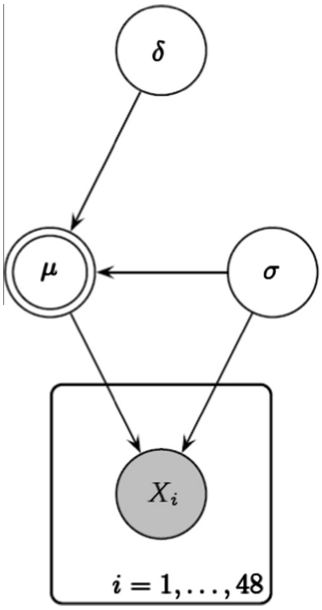
Any value, X_i, depends on its parent nodes, μ and σ^2, which must be estimated. We assumed uninformative priors for σ and δ (Gelman and Hill, 2007; Rouder et al., 2009):

$$\sigma \sim \text{Cauchy}(0,1)^+ \tag{2}$$
$$\delta \sim \text{Cauchy}(0,1).$$

The Cauchy distribution is a t-distribution with 1 degree of freedom.[10] The "+" indicates the distribution is supported on the positive semi-infinite interval, $[0, \infty)$. As in conventional t-tests, effect size can also be calculated in a Bayesian t-test. The effect size, δ, is the ratio of the mean and standard deviation:

$$\delta = \mu/\sigma. \tag{3}$$

$$\mu = \delta \times \sigma. \tag{4}$$



[9] We also conducted an analysis that first transformed the differences in d' into standardized scores. The results were almost identical with the reported results that did not initially standardize the scores. We present the unstandardized analyses because interpreting the posterior of μ is more intuitive than the z scores of μ.

[10] The reasoning behind placing a prior on δ and not μ is described by Wagenmakers et al. (2010) who states that by placing a prior on effect size, the model can be used in a broader array of situations than if a prior on the mean were used instead.

References

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1365–1376. http://dx.doi.org/10.1037/a0032188.

Atkinson, R. C., & Shiffrin, R. M. (1968). *Human memory: A proposed system and its control processes*. Oxford, England: Academic Press, pp. xi, 249, http://dx.doi.org/10.1016/S0079-7421(08)60422-3.

Aue, W. R., Criss, A. H., & Fischetti, N. (2012). Associative information in memory: Evidence from cued recall. *Journal of Memory and Language, 66*, 109–122.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*(4), 461–478. http://dx.doi.org/10.1016/j.jml.2006.08.003.

Criss, A. H. & Koop, G. J. (in press). Differentiation in episodic memory. In Raaijmakers, J., Criss, A.H., Goldstone, R., Nosofsky, R., & Steyvers, M. (Eds.), *Cognitive Modeling in Perception and Memory: A Festschrift for Richard M. Shiffrin*. Psychology Press.

Criss, A. H., Aue, W., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language, 64*, 119–132.

Criss, A. H., & Malmberg, K. J. (2008). Evidence in support of the elevated attention hypothesis of recognition memory. *Journal of Memory and Language, 59*, 331–345.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64*(4), 316–326. http://dx.doi.org/10.1016/j.jml.2011.02.003.

Criss, A. H., & Shiffrin, R. M. (2004a). Context noise and item noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review, 111*(3), 800–807. http://dx.doi.org/10.1037/0033-295X.111.3.800.

Criss, A. H., & Shiffrin, R. M. (2004b). Interactions between study task, study time, and the low-frequency hit rate advantage in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 778–786. http://dx.doi.org/10.1037/0278-7393.30.4.778.

Criss, A. H., & Shiffrin, R. M. (2005). List discrimination in associative recognition and implications for representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1199–1212. http://dx.doi.org/10.1037/0278-7393.31.6.1199.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–478. http://dx.doi.org/10.1037/0033-295X.108.2.452.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59*(3), 361–376. http://dx.doi.org/10.1016/j.jml.2008.06.007.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42*(1), 204–223. <http://www.jstor.org/stable/2958475>.

Dorfman, D., & Glanzer, M. (1988). List composition effects in lexical decision and recognition memory. *Journal of Memory and Language, 27*(6), 633–648. http://dx.doi.org/10.1016/0749-596X(88)90012-5.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1–67. http://dx.doi.org/10.1037/0033-295X.91.1.1.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: Robert E. Krieger.

Hemmer, P., & Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1947–1952. http://dx.doi.org/10.1037/t19791-000.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics, 4*(3), 227–241. http://dx.doi.org/10.1109/TSSC.1968.300117.

Jeffreys, H. (1961). *Theory of probability* (2nd ed.). Oxford: Clarendon Press.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39*(2), 348–363. http://dx.doi.org/10.3758/s13421-010-0007-6.

Koop, G. J., Criss, A. H., & Malmberg, K. J. (2015). The dynamic effects of feedback and test composition over the course of recognition memory testing. *Psychonomic Bulletin & Review, 22*, 509–516.

Koppell, S. (1977). Decision latencies in recognition memory: A signal detection theory analysis. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 445–457.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*(7), 293–300. http://dx.doi.org/10.1016/j.tics.2010.05.001.

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. http://dx.doi.org/10.1177/1745691611406925.

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. San Diego, CA, US: Elsevier Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573–603. http://dx.doi.org/10.1037/a0029146.

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 970–988. http://dx.doi.org/10.1037/a0015728.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review, 120*(1), 155–189. http://dx.doi.org/10.1037/a0030851.

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

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*(4), 335–384. http://dx.doi.org/10.1016/j.cogpsych.2008.02.004.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23*(2), 115–119. http://dx.doi.org/10.1177/0956797611430692.

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 319–331. http://dx.doi.org/10.1037/0278-7393.30.2.319.

Malmberg, K. J., Lehman, M., Annis, J., Criss, A. H., & Shiffrin, R. M. (2014). *Consequences of testing memory*. In B. Ross (Ed.). *Psychology of learning & motivation* Vol. 61. (pp. 285–313).

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 616–630. http://dx.doi.org/10.1037/0278-7393.28.4.616.

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 322–336. http://dx.doi.org/10.1037/0278-7393.31.2.322.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30*(4), 607–613. http://dx.doi.org/10.3758/BF03194962.

Mensink, G., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review, 95*(4), 434–455. http://dx.doi.org/10.1037/0033-295X.95.4.434.

Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General, 128*(4), 403–415. http://dx.doi.org/10.1037/0096-3445.128.4.403.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review, 120*(2), 356–394. http://dx.doi.org/10.1037/a0032020.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134. http://dx.doi.org/10.1037/0033-295X.88.2.93.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604. http://dx.doi.org/10.3758/BF03196750.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. http://dx.doi.org/10.3758/PBR.16.2.225.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review, 108*(1), 257–272. http://dx.doi.org/10.1037/0033-295X.108.1.257.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166. http://dx.doi.org/10.3758/BF03209391.

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. *Rational Models of Cognition*, 73–95.

Spiegelhalter, D. J., Thomas, A., & Best, N. (1996). Computation on Bayesian graphical models. *Bayesian Statistics, 5*(5), 407–425.

Turner, B. M., Dennis, S., & Van Zandt, T. (2013). Likelihood-free Bayesian analysis of memory models. *Psychological Review, 120*(3), 667–678. http://dx.doi.org/10.1037/a0032458.

Wagenmakers, E., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*(3), 158–189. http://dx.doi.org/10.1016/j.cogpsych.2009.12.001.

Wagenmakers, E., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology, 48*(3), 332–367. http://dx.doi.org/10.1016/j.cogpsych.2003.08.001.

Xu, J., & Malmberg, K. J. (2007). Moldeing the effects of verbal- and nonverbal-pair strength on associative recognition. *Memory & Cognition, 35*, 526–544.