Contents lists available at ScienceDirect

Journal of Memory and Language

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

Item effects in recognition memory for words

Emily Freeman [a], Andrew Heathcote [a,*], Kerry Chalmers [a], William Hockley [b]

[a] School of Psychology, The University of Newcastle, Australia
[b] Department of Psychology, Wilfrid Laurier University, Canada

ARTICLE INFO

Article history:
Received 2 September 2008
Revision received 17 September 2009
Available online 20 October 2009

Keywords:
Recognition memory
Word frequency mirror effect
Lexical memory
Mixed effects models
Memory models
Item effects
Lexical-decision task
Naming

ABSTRACT

We investigate the effects of word characteristics on episodic recognition memory using analyses that avoid Clark's (1973) "language-as-a-fixed-effect" fallacy. Our results demonstrate the importance of modeling word variability and show that episodic memory for words is strongly affected by item noise (Criss & Shiffrin, 2004), as measured by the orthographic similarity between experimental items. We found that the word frequency effect was not related to the item noise effects, whereas the effect of neighborhood density, which measures the similarity of a word to all other words in the lexicon, was greatly attenuated when item noise was controlled. Our results are also consistent with a likelihood based recognition decision mechanism that produces a mirror effect by taking into account item and subject characteristics.

Introduction

Memory for words is one of the most intensively studied topics in the Cognitive and Neural Sciences, both in terms of the lexical memory systems that underpin linguistic capacities, and in terms of episodic memory for verbal stimuli. Here we apply new computational and statistical techniques to the question of how lexical variables affect episodic recognition memory. We focus on two important variables for both lexical and episodic memory; word frequency (Baayen, Piepenbrock, & van Rijn, 1993) and neighborhood density (Coltheart, Davelaar, Jonasson, & Besner, 1977).

In the first part of this paper we review theories of the episodic frequency effect and explore their extension to density effects. In the following section we describe how random item effects analyses (Baayen, Davidson, & Bates, 2008) can be combined with measures of inter-item similarity (Cleary, Morris, & Langley, 2007; Yarkoni, Balota, & Yap, 2008) and estimates of other item characteristics derived from large text corpora (Baayen et al., 1993) to gain insight into the causes of both effects. We then report the results of applying this analysis to an experiment that examines word frequency and density effects using the same set of words with four groups of subjects who studied the words in different ways.

Theories of the frequency effect

Higher frequency words (i.e., words that occur more often in natural language) speed responding in lexical memory tasks, such as naming printed words and deciding if a string of letters is or is not a word (lexical decision) relative to lower frequency words (e.g., Andrews, 1997). In contrast, higher frequency words have reduced recognition memory accuracy relative to lower frequency words (e.g., Glanzer & Adams, 1985). The same is true for higher vs. lower density words (Cortese, Watson, Khanna, & McCallion, 2006; Cortese, Watson, Wang, & Fugett, 2004; Glanc & Greene, 2007; Heathcote, Ditton, & Mitchell, 2006), where density is defined as the number of words that differ from a base word by one letter (Coltheart et al., 1977) or

[*] Corresponding author. Address: The School of Psychology, Psychology Building, The University of Newcastle, University Avenue, Callaghan, 2308, Australia. Fax: +61 2 49216906.
E-mail address: andrew.heathcote@newcastle.edu.au (A. Heathcote).

0749-596X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2009.09.004

phoneme (e.g., Cleary et al., 2007). Density has been found to have a smaller effect than word frequency in episodic tasks (Heathcote et al., 2006). In the case of the lexical decision, density effects may be limited to low frequency words, and the density effect is larger when the nonwords used in the lexical-decision task are unlike words (e.g., are not pronounceable).

At least three types of explanation have been offered for the recognition memory word frequency effect. One is that lower frequency words attract more attention during study (e.g., Glanzer & Adams, 1990; Malmberg & Nelson, 2003). Malmberg and Nelson focus on "early-phase" word identification processes as the cause of attention differences, consistent with the observed pattern of slower performance in lexical memory tasks for lower frequency words. Given this hypothesis about word frequency, it seems reasonable to suggest that slower identification for low density words might cause them to receive increased study attention compared to high density words, resulting in more accurate episodic recognition.

A second suggested cause is that lower frequency words have more distinctive episodic representations. Lower frequency words are assumed to be more distinctive either because they have a lower level of context noise (e.g., Dennis & Humphreys, 2001) or a lower level of item noise (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Context noise occurs due to an increase in the strength and/or probability of chance matches between a reinstatement of the study context and other contexts in which the test item has occurred. Item noise occurs due to an increase in the strength and probability of chance matches between a test item and memory traces of studied items.

Given that it is also reasonable to expect a greater level of chance orthographic and phonological matches for high density test items, density effects might also be explained by increased item noise. The density effect might also be caused by context noise because the constituents of high density words (i.e., the parts they share with many other words) will appear in more pre-experimental contexts than the constituents of low density words. In Dennis and Humphrey's (2001) model, however, words are assumed to usually have a "unitized" representation unless the encoding task specifically emphasises their constituents. Hence, in their model, context noise is unlikely to explain the density effect.

A third type of explanation attributes episodic frequency effects for unstudied (new) and studied (old) test items to different causes. Low frequency new items are assumed to be less familiar (Joordens & Hockley, 2000; Reder, Angstadt, Cary, Erickson, & Ayers, 2002) than high frequency new items, because they have been less frequently experienced in the past, and so are less likely to cause a false alarm (i.e., an old response to a new item). However, low frequency old items are assumed to be easier to recollect (Joordens & Hockley, 2000; Reder et al., 2002). When these latter factors are sufficiently influential, they can overcome the effect of familiarity, causing a higher hit rate (i.e., an old response to an old item) for low frequency words. The combined pattern of fewer false alarms and more frequent hits for low than high frequency words is called a mirror effect. A reviewer noted that the opposite direction of frequency and density effects between lexical and episodic recognition tasks might also be called a mirror effect. We agree, but adhere to the standard reference of the term "mirror effect" to opposite effects on false alarms and hits in episodic recognition.

If familiarity is, at least in part, determined by the frequency with which the constituents of words are experienced, high density words should be more familiar than low density words, because their constituents occur more often in other words. An effect of density on recollection is also possible if low density words have more distinctive episodic representations. Reder et al. (2002) suggest that low frequency words are more distinctive because of lower context noise, that is, because they have appeared in fewer pre-experimental contexts than high frequency words. As in Dennis and Humphreys' (2001) model, whether this context noise mechanism could explain the density effect depends on whether context noise is associated with the constituents of words. However, Yonelinas (2002) suggests that recollection is mainly affected by semantic properties, and so in his theory there is unlikely to be any relationship between density and recollection.

The mirror effect

As previously described, two process theories assume that the word frequency mirror effect in recognition memory arises from the opposing effects of two underlying factors. Theories which attribute the word frequency effect to a single factor can also produce a mirror effect by assuming different decision criteria for low and high frequency items (e.g., Gillund & Shiffrin, 1984). More recent single-factor theories predict a mirror effect because recognition decisions are based on an estimate of the *likelihood* that a test item is old. These theories include REM (Retrieving Effectively from Memory, Shiffrin & Steyvers, 1997), SLiM (Subjective Likelihood Model, McClelland & Chappell, 1998), BCDMEM (Bind-Cue-Decide Memory Model, Dennis & Humphreys, 2001) and ALT (Attention-Likelihood Theory, Glanzer & Adams, 1990).

In these theories, likelihood is estimated by combining a test item's match to episodic memory (i.e., memory strength) with knowledge about items. In ALT, the knowledge is specific to the particular test item (i.e., the level of attention given to that item during study). In the other likelihood theories this knowledge is more general. In REM, item knowledge is captured by the *g* parameter, which is based on all items previously experienced. In SLiM, item knowledge is captured by the *p* parameter, which is based on all items in the study list. In all cases, the outcome is that likelihood values for new and old test items of the same type (e.g., frequency) tend to be centered around a common origin (see Glanzer, Adams, & Iverson, 1991, for discussion related to ALT). Centering spreads the effect of the single factor to both new and old items, and so tends to create a mirror effect.

Likelihood theories naturally provide an explanation of the generality of the mirror effect, which occurs for a range of item variables besides word frequency (Glanzer & Adams, 1985). This is the case because any single item factor that influences accuracy will tend to have its effect spread to both new and old items due to centering. To the degree that likelihood theories predict a mirror effect due to any item property that affects accuracy, they will predict that density causes a mirror effect.

Because item mirror effects tend to be centered on a common origin, likelihood theories also predict approximately equal response bias for each type of item (e.g., for all combinations of low and high density and frequency items). That is, the tendency to respond "old" averaged over new and old items is the same for each item type. The single-factor criterion shift explanation, and two process explanations, can also accommodate equal bias, but do not predict this finding unless further elaborated to explain why the appropriate criterion shift, or tradeoff between the effects of the two processes, occurs.

In summary, most of the factors that have been identified as causing the word frequency mirror effect in episodic recognition might plausibly also apply to density effects. Two of these factors are specific to old items, differences in attention during study and differences in the probability of recollection, whereas two other factors, differences in distinctiveness and familiarity, apply to both new and old items. A mirror effect can occur either because of a tradeoff between two factors (e.g., familiarity and recollection) or because of the effect of a single factor on the recognition decision process (e.g., due to a criterion shift or likelihood based decisions). In the next section we discuss new statistical analyses that can be used to investigate the causes of the word frequency and density effects.

Random item effects models

Although word frequency and density effects on episodic recognition accuracy are reliable in the sense of being experimentally replicable, reported findings rely on "set-level analyses" (Lamberts, Brockdorff, & Heit, 2003) that aggregate over items. Rouder and Lu (2005), see also Rouder, Lu, Morey, Sun, and Speckman (2008) showed that such set-level analyses produce biased estimates of the signal detection theory d' measure of accuracy (Macmillan & Creelman, 2005). Their application of signal detection theory assumed that subjects classify a test item as old if its episodic memory strength is greater than a criterion (c). New and old test item memory strengths were assumed to have equal variance normal distributions that differ in their means by an amount equal to d'. In a set-level analysis, item variation is confounded with memory strength variation, which causes d' (i.e., the difference between the means of memory strength distributions) to be underestimated.

The bias identified by Rouder and Lu (2005) is asymptotic, meaning that it cannot be ruled out by experimental replication or using large item sets. Where the level of item variation differs between experimental conditions (e.g., comparisons of high vs. low frequency or density words) different levels of underestimation can confound effect tests. Fortunately, the bias can be avoided by treating items as a random effect. In the following sections we review the theory and practice of random effect modeling and discuss the advantages which this approach brings to the analysis of recognition memory for words.

Random-effects-models

Experimental factors can be treated as arising from effects that are either fixed or random. Fixed effects are associated with experimental manipulations that are exactly repeatable. For example, a "test-type" factor, with levels consisting of test items that either have or have not been studied (i.e., old vs. new), is a fixed effect. Random effects are most familiar to psychologists when applied to subjects, as the standard assumption made in analysis of variance (ANOVA) and covariance (ANCOVA) is that subjects are a random effect. This assumption is made because interest usually focuses on what a sample of subjects tells the experimenter about the population of subjects. Clark (1973) argued that a population focus is also appropriate for linguistic items as experimenters usually select item sets to be representative of a population of items (e.g., high frequency words). In this case it is appropriate to also treat items as a random effect, and Clark characterized the assumption that items are a fixed effect as the "language-as-fixed-effect-fallacy".

Often experiments with random factors aim to compare the characteristics of different populations, such as high vs. low frequency words or male vs. female subjects. The different populations constitute a fixed effect (e.g., the effect of word frequency or gender) whereas the members of the population sampled in an experiment (i.e., the particular words or subjects) constitute the random effect. Populations are usually modeled by parametric distributions, such as the normal distribution. The parameters of the population distributions, such as their means and standard deviations, and potentially other parameters such as correlations in the case of multivariate population distributions, can be used to describe the populations.

Parameters that describe the nature of the random variability in a population, such as standard deviations and correlations, are usually called random effect parameters. The population means are usually called fixed effect parameters because they can be separated from the random effects. For example, a normal distribution with mean (μ) and standard deviation (σ), $N(\mu, \sigma)$, can be written as the sum of the mean and a zero centered random effect: $\mu + N(0, \sigma)$. However, it is important to note that a mean estimated by a fixed effects analysis is not necessarily the same as a mean estimated by a random effects analysis. This is the case because random-effects-models obtain estimates of population parameters, and associated measures of estimation uncertainty (e.g., standard errors), in a way that takes into account both the assumed form of the population distribution and the fact that a new sample from the population will likely differ from the one presently under consideration. As a result, statistical tests based on random effects analysis can be used to make valid inferences about the population.

In contrast, when an effect is treated as fixed, inferential results are valid for the particular effect levels (e.g., items or subjects) used in the experiment, but underestimate error in generalizing to another sample (e.g., estimated confidence intervals are too narrow). Hence, assuming an effect is fixed when it is actually random produces overly optimistic estimates of the reliability of the effect in a new sample, resulting in inflated Type 1 error rates. Rouder and Lu (2005) showed that erroneously treating either subject or item effects as fixed in a recognition memory paradigm can cause an inflation of Type 1 error rates that can be surprisingly large, as well as decreasing power (see also Baayen et al., 2008). Hence, statistical analysis should ideally treat both subjects and items as random effects in the same analysis.

Fitting random-effects-models

An analysis with simultaneous random item and random subject effects is technically difficult in designs used to examine episodic recognition memory. Such designs are necessarily incomplete with respect to the crossing of items and subjects, as items are divided into a studied (old) and unstudied (new) set for each subject. That is, any one subject only experiences a given item as old or new, not both. As Lamberts et al. (2003) note, this design limitation is difficult to avoid, as test items cannot be repeated within subjects without changing the nature of the task to temporal or list-context discrimination.

Fortunately, recently developed Bayesian (Rouder et al., 2007) and maximum likelihood (Bates, 2005) analyses are applicable to incomplete designs, as long as: (a) random item and subject effects are assumed to be additive and (b) a large and variable sample of subjects and items is available. The latter assumption held for our large data set consisting of over 27,000 observations from around 100 subjects and 300 items. We applied maximum likelihood analysis to our data largely because of convenience; it allowed us to efficiently specify and compare a large range of random effects models. In particular our analysis used the linear mixed effects package (*lme4*, Version 0.99875-9) for the R statistical environment (R Development Core Team, 2007). This package can fit generalized linear mixed models (McCullagh & Nelder, 1989), one of which, the binomial probit model, is equivalent to the signal detection model examined by Rouder et al. (2007, 2008) and Rouder and Lu (2005).

Mixed effect models (i.e., models with a mixture of fixed and random effects) take advantage of the separation between fixed and random effects parameters by allowing for different models of experimental factors for each type of parameter. For example, it might be assumed that fixed effects vary as a function of all of the factors manipulated in an experimental design, whereas random effects depend on only a subset of these factors. Usually the random effects model is simpler than the fixed effects model because an overly complex random effects model can make estimation numerically unstable (i.e., it either fails or produces unreasonable parameter estimates). In contrast, specifying a complex fixed effects model that can exactly fit the means in each experimental condition (i.e., a "saturated" model, the conventional assumption in ANOVA) does not usually cause estimation problems. Consequently a saturated fixed effects model is usually assumed in mixed model analysis. In contrast, a forward model selection strategy (i.e., comparing increasingly complex models) must be adopted to identify a random effects model that adequately describes the data with a minimum number of parameters.

For example, as a first step in forward selection, models with only random subject effects or only random item effects might be compared to a more complex model with additive item and subject effects. If the latter model is preferred it indicates that the data contain reliable item and subject random effects. This model would then be adopted and compared to more complex models incorporating effects of experimental manipulations on random effect parameters, and the process iterated until the best model is found (Baayen et al., 2008). A range of criteria may be used to guide the model selection process, such as a significant improvement in fit or a reduction in information criteria which balance misfit with a penalty for the number of model parameters. Two such measures, AIC, the Akaike Information Criterion and BIC, the Bayesian Information Criterion (Myung & Pitt, 1997; Vaida & Blanchard, 2005), are provided by the *lme4* software.

BIC imposes the harshest penalty for complexity, and we found it generally performed best in terms of selecting models with numerically stable random effect parameter estimates with reasonable values. We report model selection results in terms of the probability that a given model is best within a set of models, which can be obtained by a transformation of BIC (Wagenmakers & Farrell, 2004). We label this probability p_{BIC}; it can be directly interpreted as the probability that a model is correct (Raftery, 1995). We assess the strength of evidence associated with a p_{BIC} value against conventions similar to those suggested by Raftery: $p_{\mathrm{BIC}} < .75$ provides weak evidence, $0.75 \leqslant p_{\mathrm{BIC}} < .95$ provides positive evidence, $0.95 \leqslant p_{\mathrm{BIC}} < .99$ provides strong evidence and $p_{\mathrm{BIC}} > .99$ provides very strong evidence.

Random effect parameters are often treated as "nuisance" parameters that must be estimated for statistical reasons but which have no psychological interest. However, this is not the case in our application because random effect parameters provide a test of predictions made by likelihood theories of episodic recognition. These theories predict that study items and subjects with low values in the new condition should have larger values in the old condition (i.e., a mirror effect). To illustrate, suppose subjects or items A and B have estimates of 1 and 2, respectively, in the new condition, and estimates of 4 and 3, respectively, in the old condition. The mirroring of orders results in a negative correlation between estimates from the new and old condition. As a result, model selection should support the inclusion of a test-type factor (i.e., a factor that differentiates studied and unstudied items) in the random effects model, and the associated correlation parameter between the levels of this factor should be negative. Rouder et al. (2007) found trends consistent with the latter prediction, but they were not statistically reliable. Our experiment, with over four times more observations, provides a more powerful test of both predictions.

Random parameter estimates associated with the test-type factor may also help to identify the mechanism that mediates an often replicated finding based on set-level analysis of episodic recognition memory. A plot of hit rates against false alarm rates, a Receiver Operating Characteristic (ROC), has a slope less than one when the rate measures are z-transformed (i.e., a zROC). In a signal detection theory framework this finding indicates that memory strength for old test items is more variable than memory strength for new test items. Several causes have been suggested for this finding, including causes related to underlying memory mechanisms (e.g., variations in recollection, Yonelinas, 1994, or study attention, Ratcliff, Sheu, & Gronlund, 1992, for old test items) and causes related to the decision process when zROC measurement is based on confidence ratings (e.g., variation in confidence criteria, Mueller & Weidemann, 2008, or dynamic decision processes, Ratcliff & Starns, 2009). As a set-level analysis confounds item and memory strength variation it cannot determine if these causes have their effect through an item specific mechanism. If this is the case, estimates of old test item variance will be greater than estimates of new test item variance in our analysis.

Item covariates

Random item effects models provide another important advantage; they enabled us to control issues related to the particular set of items used in an experiment, through the use of item covariates. We used estimates of word frequency and density derived from the CELEX corpus (Baayen et al., 1993) to select item sets instantiating a tightly controlled 2 × 2 (high–low) factorial manipulation of these variables. We adopted this high vs. low item set design because it is conventional in the analysis of frequency effects, which allowed us to compare our results to previous research. However, because we used relatively large item sets (75 items of each type) it was not possible to exactly equate item sets on variables such as average letter and bigram frequency that have been shown to be influential in lexical and episodic memory tasks (Andrews, 1997; Malmberg, Steyvers, Stephens, & Shiffrin, 2002). To control these potential confounds we used these variables as item covariates in our analyses.

We also used item covariates to examine a second issue related to the particular sample of items used by a researcher; differences in average ‘‘inter-item similarity” between item sets (Cleary et al., 2007). Cleary et al. showed that inter-item similarity affected episodic recognition memory for nonwords; we investigate whether there is also an effect on memory for words. If there is a reliable relationship between our inter-item similarity covariates and memory performance it will support a role for item noise in episodic recognition. Inter-item similarity effects may also provide an item noise explanation of the effects of frequency and density. By definition, high density words are more likely to have higher inter-item similarity than low density words. It has also been suggested that high frequency words contain more common spelling and sound patterns than low frequency words (Estes & Maddox, 2002; Landauer & Streeter, 1973), and so they may also have higher inter-item similarity. In both cases, higher inter-item similarity could cause higher item noise, and hence reduced accuracy. If a reliable item set effect is attenuated or disappears when inter-item similarity measures are used as covariates, item noise is implicated as a cause of the item set effect.

Inter-item similarity

Because there is no agreed measure of similarity between words we considered five possible metrics of inter-item similarity: phonologically and orthographically based stem and Levenshtein similarity, and orthographic neighbor overlap. Both orthographic and phonological variants were used as similarity amongst words might depend on their constituent letters and/or their constituent phonemes. The stem similarity metrics provide a binary measure of similarity, classifying an item pair as either mismatching or matching. A pair matches if their first or last three phonemes, as used by Cleary et al. (2007), or letters are the same. Following Cleary et al. we quantified stem similarity for an item as equal to the sum of its matches with all other items in the experimental set.

The Levenshtein measure quantifies similarity between pairs of items in a more continuous manner than stem similarity. It is the inverse of Levenshtein distance, a measure of dissimilarity equal to the number of deletions, additions and substitutions required to transform one string of letters into another. Hence, Levenshtein similarity indicates that two strings are similar if it takes only a few such operations to convert one into the other. Levenshtein distance has found wide application in computer science (e.g., it can be used by spell checkers to guess potential corrections for misspelled words). Yarkoni et al. (2008) also demonstrated the psycholinguistic utility of orthographic Levenshtein distance by showing it to be a good predictor of performance in naming and lexical-decision tasks. We used the inverse of Levenshtein distance (i.e., a measure of *similarity*) as the other covariates measured similarity. This enabled easier comparison among covariate effects (other methods of converting Levenshtein distance to a similarity measure, such as multiplying distance by minus one, did not produce appreciably different results).

Compared to stem similarity, the Levenshtein measure also takes a different approach to the ‘‘alignment problem” (i.e., how to compute similarity between items of different lengths). Stem similarity aligns items on their first or last three constituents, ignoring any length difference, whereas the Levenshtein measure takes account of length differences. For example, our item sets consisted of words varying between four and six letters. At a minimum, two additions are required to transform a four letter word to a six letter word, which reduces Levenshtein similarity. In contrast stem similarity counts a four letter word as matching a six letter word as long as they share either their first or last three constituents.

Our final inter-item similarity covariate, neighbor overlap, is the number of times that words in an item's orthographic neighborhood are the same as, or are contained in, the neighborhoods of other experimental items. Although this covariate has not been used in previous work it was natural to include it here given that we examined the effects of a manipulation of orthographic neighborhood density. All else being equal, it is likely that neighbor overlap will confound our density manipulation because our high density items are likely to have higher neighborhood overlap than our low density items.

Item covariate model selection

We used a relatively large range of item covariates in order to perform a thorough investigation of potential confounding causes of the word frequency and density effects. However, this approach may be compromised by the statistical issue of "over-fitting". That is, when all seven covariates are used, item-set effects may be attenuated due to chance rather than systematic relationships. One approach to this issue is to use only a subset of covariates that have a reliable effect by some criterion. Because of correlations amongst covariates it is not sufficient to identify this subset by applying the criterion to the results of an initial fit in which all covariates are included. Instead, we fit and compared all 128 (i.e., 2^7) possible covariate subset models, ranging from a model with no covariates to the model with all covariates, and all combinations in between. Each model was specified separately and fit by *lme4* (we wrote programs in R to automate the process as the computing time required to fit all of these models was substantial). We used BIC as our model selection criterion and report our results in terms of p_{BIC}.

Subset selection has a further weakness; it assumes there is no model uncertainty. That is, it assumes that the correct subset model can be identified with certainty, which is unlikely. However, having fit all subset models and obtained their p_{BIC} values, it is not necessary to select only one model. Instead, we can examine the average word frequency and density effects predicted by a weighted average of all models, where the weight is given by the model's p_{BIC} value. This procedure is called Bayesian Model Averaging (BMA) and it has been shown to have better predictive validity than assuming any single model (Raftery, 1995).

We were also interested in identifying which item covariates have a systematic association with episodic recognition memory performance and the size of their effects. Once again BMA provides a solution that takes account of model uncertainty. The probability that each covariate has an association can be obtained by summing model probabilities over the models in which it occurs. The estimated effect of each covariate, assuming that is does have an effect (i.e., should be included in a covariate model) is the probability weighted average of estimates from the models in which it occurs. Standard errors for these average estimates tend to be wider than for estimates from any individual containing the covariate (Raftery, 1995) as they take proper account of model uncertainty amongst the models that contain the covariate (in our case 64 models).

Experiment

We investigated word frequency and density effects using a 2 × 2 high-low item-set design and analyzed our results using additive random item and subject effects mixed models. Model selection analyses were used to determine an appropriate random effects model and to investigate the role of item covariates. These analyses aimed to answer a number of psychological questions:

(1) Do subject and item variation both have an appreciable effect on recognition memory performance?
(2) Does the test-type factor enter into the random effects model? If so, are old condition effects more variable than new condition effects, explaining set-level zROC results, and are new and study (i.e., old–new) effects negatively correlated, as predicted by likelihood theories?
(3) Are the word frequency and density effects previously found with set-level analysis reliable when random item effects are taken into account, or are they Type 1 errors resulting from the overestimation of reliability by set-level analysis?
(4) Do word frequency and/or density effects enter into the random effects model? If so, can their effects on accuracy be explained by differences in variability between high and low sets rather than differences in mean memory strength?
(5) Is episodic recognition influenced by inter-item similarity covariates? If so, is the direction of the effect as predicted by item-noise theories; a decrease in accuracy as inter-item similarity, and hence item noise, increases?
(6) Can item noise, as measured by inter-item similarity covariates, or effects of the frequency of word constituents, such as letters or bigrams, explain the effects of word frequency or density?

The experiment that we report had four study conditions manipulated between-subjects: two "free-study" conditions, where subjects were told only to memorize words studied for 1.5 s each for a later test; a lexical-decision task; and a naming task. In one of the free-study conditions words-only were studied. In the other, both words and nonwords (the same as used in the lexical-decision and naming tasks) were studied. The word-only condition is most typical of previous experiments examining word frequency and density effects. The word–nonword condition provides a control to examine the effects of including nonwords during study, and to provide an appropriate comparison with the other two tasks, which also included nonwords in the study task.

The variety of tasks we examined allows us to determine the robustness of item-set effects across different encoding manipulations. Note, however, that the study conditions differ in study-test lags and number of study items hence it is not appropriate to make comparisons of the overall level of performance across these conditions. The naming study task and the lexical-decision study task also yielded data that provided a manipulation check for our item sets in terms of lexical memory performance. As these study tasks also focus encoding specifically on identification processing they allow us to examine a prediction made by Malmberg and Nelson's (2003) early-phase attention hypothesis; an effect of word frequency on recognition memory should only occur if there is a corresponding effect on study task performance. The same prediction is made for recognition memory density effects if they are also caused by attention differences in early-phase identification processes. In order to address these issues, in our results section we first report the outcomes of an analysis of performance in the lexical tasks before reporting results related to episodic recognition.

Method

Subjects

We tested 111 University of Newcastle undergraduates (31, 25, 28 and 27 in the word-only, word–nonword, lexical-decision and naming conditions, respectively). Subjects with accuracy less than 60% in the recognition task, less than 85% in lexical-decision, or who failed to respond on more than 10% of study or test trials were removed in the final sample (7, 2, 3, and 7 in the word-only, word–nonword, lexical-decision and naming conditions, respectively).

Procedure

Subjects were presented with lists of words (word-only condition) or words and nonwords (word–nonword, lexical-decision and naming conditions) and asked to remember the words. In the word-only and word–nonword conditions study items were presented for 1.5 s. In the lexical-decision condition, subjects indicated whether each item was a word or a nonword, by pressing either the left (labeled *word*) or right (labeled *nonword*) outermost button on a 6-button response pad. The next study item appeared immediately after the lexical decision response was given. In the naming condition, subjects were asked to name each item aloud, and items remained on screen for 3 s. Naming time was recorded by a voice key. In the word–nonword, lexical-decision and naming conditions study lists were 30 words and 30 nonwords. In the word-only condition study lists were 30 experimental words, and 8 primacy and 8 recency buffer words. At test subjects used a 6-button response pad to make old/new decisions on 60 of each type using a 6-point confidence scale (1 = *sure-old* to 6 = *sure-new*). A maximum of 5 s was allowed for each test response, and the next test trial started immediately after a response was made. Subjects completed five study/test cycles. Assignment of items to lists, new/old, and presentation order, was randomly determined for each subject.

Items and item covariates

Items consisted of 300 words, 75 in each set making up a factorial combination of high and low density and frequency, and 300 nonwords, with equal numbers of 4, 5 and 6 letter items in each set (see Appendix). Table 1 gives the average density, frequency and covariate values for each word set, and overall correlations between these variables, and standard deviations for each. In most cases the magnitudes of the correlations are relatively small, indicating that the measures address different item characteristics.

Inter-item similarity covariates were calculated relative to test items (words) only, as phonology was not available for half of the nonwords, which were not pronounceable. The stem measures give the average number of matches out of a possible 299. The Levenshtein measures are the inverse of the mean distance between each item and the other 299 items. Hence, the average overall Levenshtein similarity of 0.22 corresponds to a distance of 4.5 (i.e., an average of 4.5 additions, deletions and substitutions to transform one word into another). Neighbor overlap is the average number of times an item or its neighbors matches the other 299 items or their neighbors.

Results

Because mixed model analysis provides a rich set of outcomes we first provide an overview of the different types of results. Results are presented graphically, first for study tasks (Fig. 1) and then for episodic recognition (Figs. 2–5). As the majority of the results concern signal detection theory measures of episodic performance we overview Figs. 2–5 first. Study task performance was measured by response time (RT), which requires some different analyses than episodic recognition. We discuss these

Table 1

Item set means for low density (LN), high density (HN), low frequency (LF) and high frequency (HF), overall item standard deviations (*SD*), and correlations. Correlations given in bold are significant $p < .01$, those in italics are $.01 \leqslant p < .05$. Phon.: inter-item phonological similarity. Orth.: inter-item orthographic similarity.

	Item set means				Overall	Overall correlations							
	LNLF	LNHF	HNLF	HNHF	*SD*	F	LF	BF	SP	SO	LP	LO	NO
Density (N)[a]	2.07	2.15	9.15	9.09	4.04	.01	.24	**.44**	**.38**	**.28**	**.27**	**.34**	**.61**
$\log_{10}$ (Word frequency) (F)[a]	0.73	1.75	0.74	1.76	0.54		.06	.07	.05	.00	*.12*	.07	-.04
Letter frequency (LF)[b]	0.069	0.072	0.078	0.079	0.016			**.32**	**.23**	.06	*.12*	**.24**	**.16**
Bigram frequency (BF)[c]	1170	1288	1931	2095	1008				**.18**	**.39**	*.14*	.08	**.30**
Stem phon. (SP)[d]	0.88	0.84	1.77	2.24	1.96					**.34**	.06	*.13*	**.43**
Stem orth. (SO)	1.53	1.37	2.43	2.53	2.49						.05	*-.15*	**.46**
Levenshtein Phon. (LP)[d,e]	0.21	0.22	0.22	0.23	0.018							**.48**	.09
Levenshtein orth. (LO)[e]	0.21	0.21	0.22	0.22	0.014								*.12*
Neighbor overlap (NO)[a]	0.35	0.63	6.79	5.55	5.58								–

[a] Calculated using the N-Watch program (Davis, 2005).

[b] Malmberg et al.'s (2002) procedure was used obtain mean token (i.e., word frequency weighted) letter frequency for each word.

[c] Token (i.e., word frequency weighted) bigram frequencies taken from the N-Watch program (Davis, 2005).

[d] Phonology for all items was obtained from the ARC nonword database (Rastle, Harrington, & Coltheart, 2002).

[e] Levenshtein distance was calculated using the "sdists" function available in the "cba" package (Buchta & Hahsler, 2007) for R (R Development Core Team, 2007) with a weight vector (1, 0, 1), corresponding to the algorithm used by Yarkoni et al. (2008).

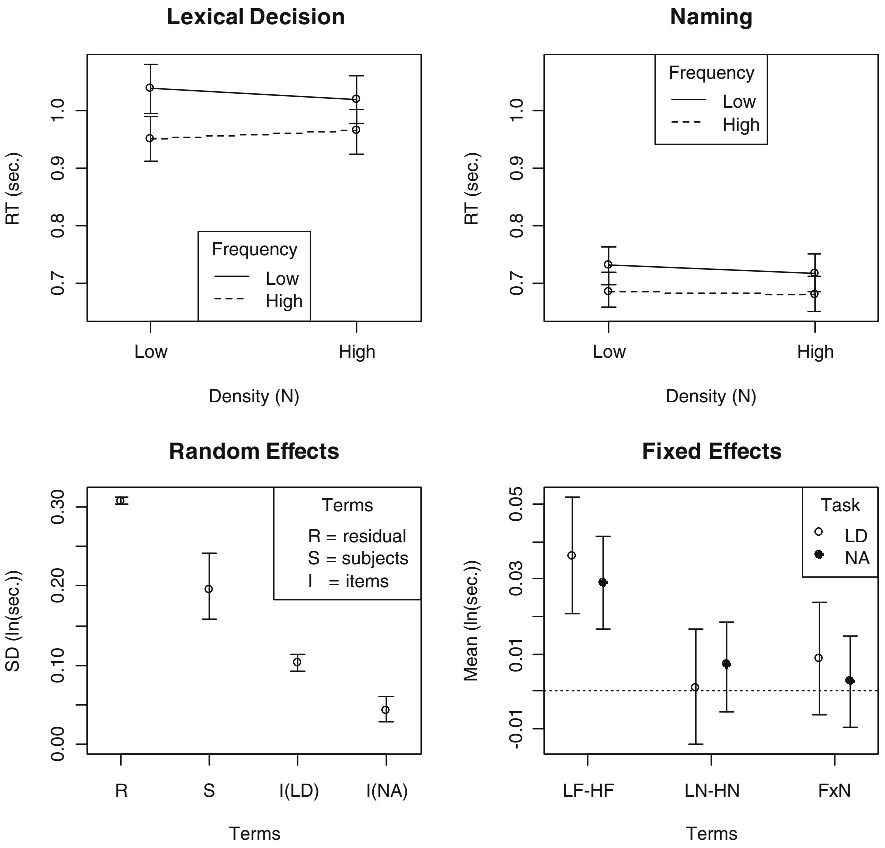

Fig. 1. Top row: Fixed effect estimates and Highest Posterior Density (HPD) 68% intervals (corresponding approximately to a normal standard error) on the seconds scale. Bottom row: Effect estimates and 95% HPD intervals on a logarithmic scale (where these effects are reliable at the $p < .05$ level intervals do not include zero). Note that LD = lexical decision, NA = naming, LF = low frequency, HF = high frequency, LN = low density, HN = high density and the FxN is the frequency by density interaction (high density minus low density frequency effect, or equivalently high frequency minus low frequency density effect). HPD intervals were obtained using the languageR package (Baayen, 2008).

differences in the following section that addresses Fig. 1 and study task results in detail.

Episodic recognition results are first presented in terms of estimates of population hit and false alarm rates and in terms of corresponding accuracy and bias estimates (Fig. 2). Population estimates were obtained by fitting a single mixed model directly to the binary test responses (i.e., old vs. new) in all study conditions. Results in Fig. 2 are accompanied by standard errors appropriate for population inference. Accuracy is indicated in terms of the d' measure and bias in terms of the c (criterion) measures of use in signal detection theory. When $c = 0$ responding is unbiased, when it is less than zero there is a bias to respond “old” and when it is greater than zero there is a bias to respond “new”. Both d' and c measures are on the inverse cumulative normal (z) scale. Hit rates and accuracy are shown in the upper portion of each panel in Fig. 2 and false alarm rates and bias are shown in the lower portion of each panel.

We also use figures to present the results of single degree of freedom word frequency and density effect and covariate effect tests and corresponding effect estimates for the d' and c measures. We use the term “reliable” to indicate effects that are significant at the two-sided $p < .05$ level. Population effect reliability can be directly apprehended from the figures by determining whether the accompanying 95% confidence intervals include a zero effect, indicated by a horizontal dotted line. These intervals were obtained from the *lme4* software by multiplying the square root of the main diagonal of the parameter variance–covariance matrix by 1.96. Parameters which correspond to the effects shown in the figures were obtained by specifying appropriate contrasts for the factors used in the analysis. Covariate effect estimates (Fig. 3) are accompanied by estimates of the probability that each covariate has an effect as estimated by BMA analysis. Discussion of the covariate model selection process accompanies discussion of Fig. 3. Word frequency and density effect estimates (Fig. 4) are shown for three types of covariate models: (a) with no covariate adjustment, (b) with adjustment for all covariates and (c) adjusted by a weighted average of covariates with weights determined by BMA. Comparison between these different effect estimates illustrates the effects of covariates on word frequency and density effects.

Estimates of random effect standard deviations and correlations parameter estimates (also on the z scale) are given graphically with 95% confidence intervals in Fig. 5, which shows estimates for the model with no covariates

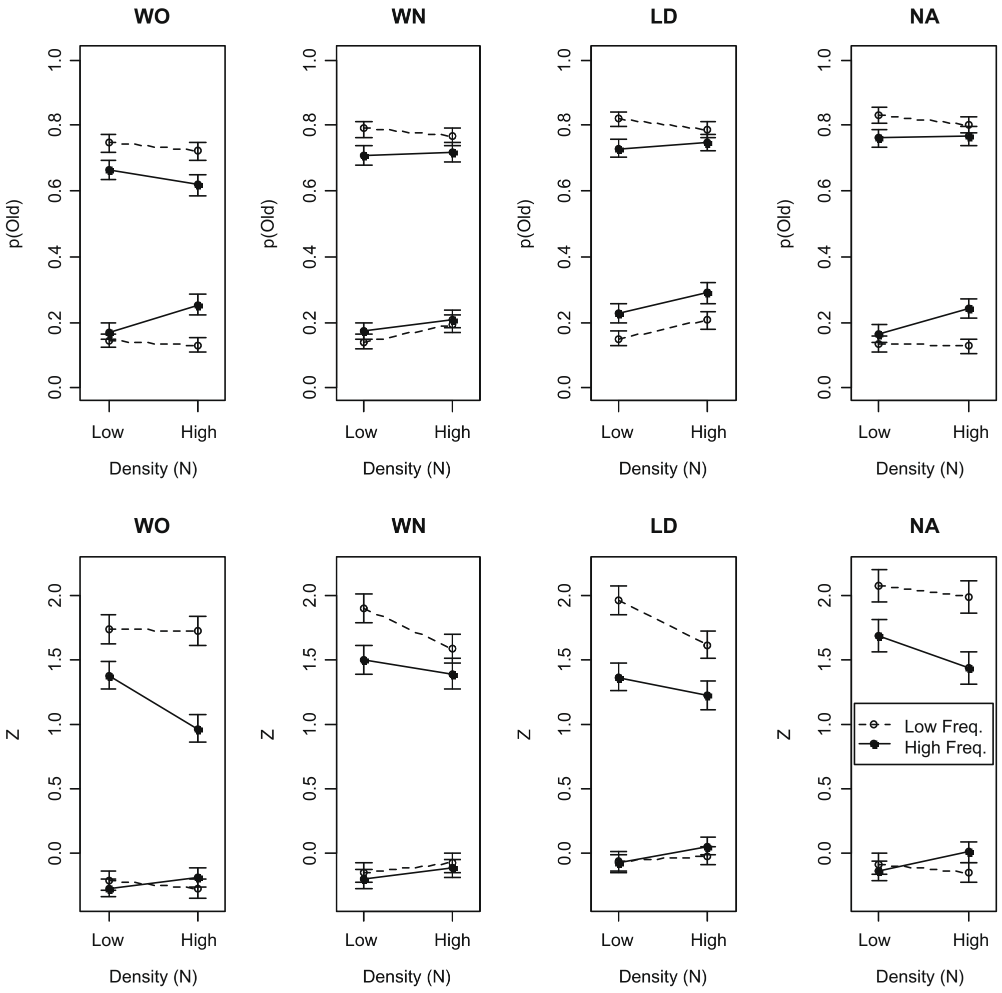

Fig. 2. Recognition memory population estimates with standard error bars. Upper row: probability of an old response, *p*(Old), for new (lower lines and symbols) and old (upper lines and symbols) test items. Lower row: accuracy (*d*′, upper lines and symbols) and bias (*c*, lower lines and symbols). Columns correspond to study conditions: WO = words-only, WN = words and nonwords, LD = lexical decision and NA = naming.

and all covariates. The confidence intervals were obtained by bootstrap methods appropriate for population level inference about the random effect parameters. That is, parameter estimates from a model fit to data were used to simulate 500 data sets. Confidence intervals were estimated based on variation amongst parameter estimates from these fits. Results of the random effects model selection process accompany discussion of Fig. 5. The random effects model shown in Fig. 5 was used to obtain results presented in Figs. 2–4.

Study tasks

For study tasks we focus on RT as accuracy was at ceiling for word naming and accuracy was high and reflected the same trends as RT for lexical decision. The *lme4* package was used to fit a mixed model with additive random item and subject effects to the RT data from both tasks simultaneously. Covariates had negligible effects and so were not included in these analyses. Following Baayen et al. (2008) we first fit a constant variance Gaussian model of RT variability using the restricted maximum likelihood method. For our data these distribution assumptions were clearly violated. Residuals were positively skewed and increased with fixed effects, consistent with general tendencies in most RT data (Wagenmakers & Brown, 2007). Hence, we report results for restricted maximum likelihood fits to the natural logarithm of RT (see Baayen, 2008), as analyses of RT data on a logarithmic scale largely remedied the distribution misspecification.

Random effects for both items and subjects received strong support over either alone ($p_{\mathrm{BIC}} > .99$). A difference in item variance between the two tasks also received strong support compared to a subject difference in task variance, or both ($p_{\mathrm{BIC}} > .99$). We could find no other factors or combinations of factors that further decreased BIC when included in either random effect. The estimated residual *SD* was larger than the subject *SD*, which in turn was larger than the item SDs (see Fig. 1). The correlation between task random effects was negligible, but the lexical-decision *SD* was reliably greater than the naming *SD* (the difference

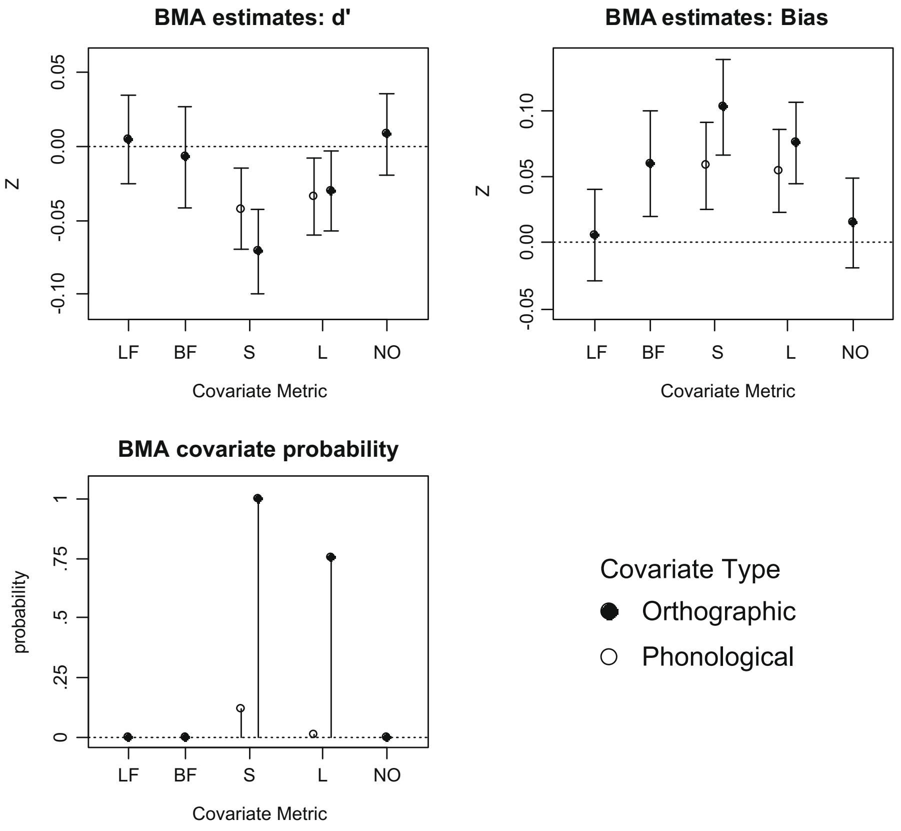

Fig. 3. Covariate estimates with 95% confidence intervals and BMA (Bayesian Model Average) covariate probabilities. Note that covariates were standardized, so the estimates represent the effect of a one standard deviation change. Reliable effects (at the $p < .05$ level) are indicated by intervals that do not cross the dotted line indicating a zero effect. Covariates metrics: S = stem and L = Levenshtein similarity, NO = neighborhood overlap, LF = letter and BF = bigram frequency.

had a 0.046–0.063 95% interval[1]). As greater variance on the log(RT) scale corresponds to greater skew on the RT scale, this task difference in *SD* is consistent with Andrews and Heathcote's (2001) finding of greater RT distribution skew in lexical-decision than naming.

Comparison of the fixed effect estimates given in the top two panels of Fig. 1 indicates that, overall, naming was much faster than lexical decision (0.05–0.17 s, 95% interval for the difference). The lower right panel of Fig. 1 provides 95% intervals for the main effects of word frequency and density and their interaction. RT was reliably greater for low than high frequency items in both tasks. In contrast, the density main effect, and its interaction with frequency, was negligible for both tasks.

The lack of density effects replicates Heathcote et al.'s (2006) finding for performance in lexical decision when it is used as a study task in an episodic recognition experiment. However, it is in contrast to the reliable density effects, at least for low frequency words, that are observed when these tasks are not followed by an episodic recognition test (Andrews, 1997). Note that the present experiment had a larger density manipulation and less word-like nonwords than in Heathcote et al. both factors would be expected to produce a larger lexical-decision density effect, but did not.

The study task results are consistent with low frequency words requiring more attention during identification than high frequency words, but not with any identification processing differences for low compared to high density words. Hence, Malmberg and Nelson's (2003) early-phase attention hypothesis can explain frequency but not density effects in episodic recognition. The cause of the apparently replicable difference in density effects between purely lexical and mixed lexical and episodic tasks is beyond the scope of the present work. We note that Heathcote et al. (2006) found reliable density effects in episodic recognition despite

[1] The *mcmcsamp* routine supplied with *lme4* can be used to perform Markov Chain Monte Carlo (MCMC) sampling from a fitted model in order to determine a Bayesian estimate of the error in mixed model parameter estimates. Uncertainty in estimates, which in the Bayesian context is called a 95% Highest Posterior Density (HPD) interval, is indicated by the difference between 2.5% and 97.5% quantiles of the posterior samples. MCMC sampling must be used with fixed effect parameters for mixed models estimated by restricted maximum likelihood (Baayen et al., 2008), but is not necessary for fixed effect parameters estimated by maximum likelihood, as was the case for our binomial probit mixed models of episodic recognition analyses. The *lme4* package does not provide any direct estimates of intervals for random effects for models estimated in either way, but *mcmcsamp* can be used for this purpose. However, we found this approach tended to produce slightly narrow interval estimates for population inference about random effects, so we used the much more computationally intensive bootstrap method for the episodic data, as inference about random effect parameters was critical in this case. For the RT data where this was not the case, we used the MCMC method to obtain HDP intervals.

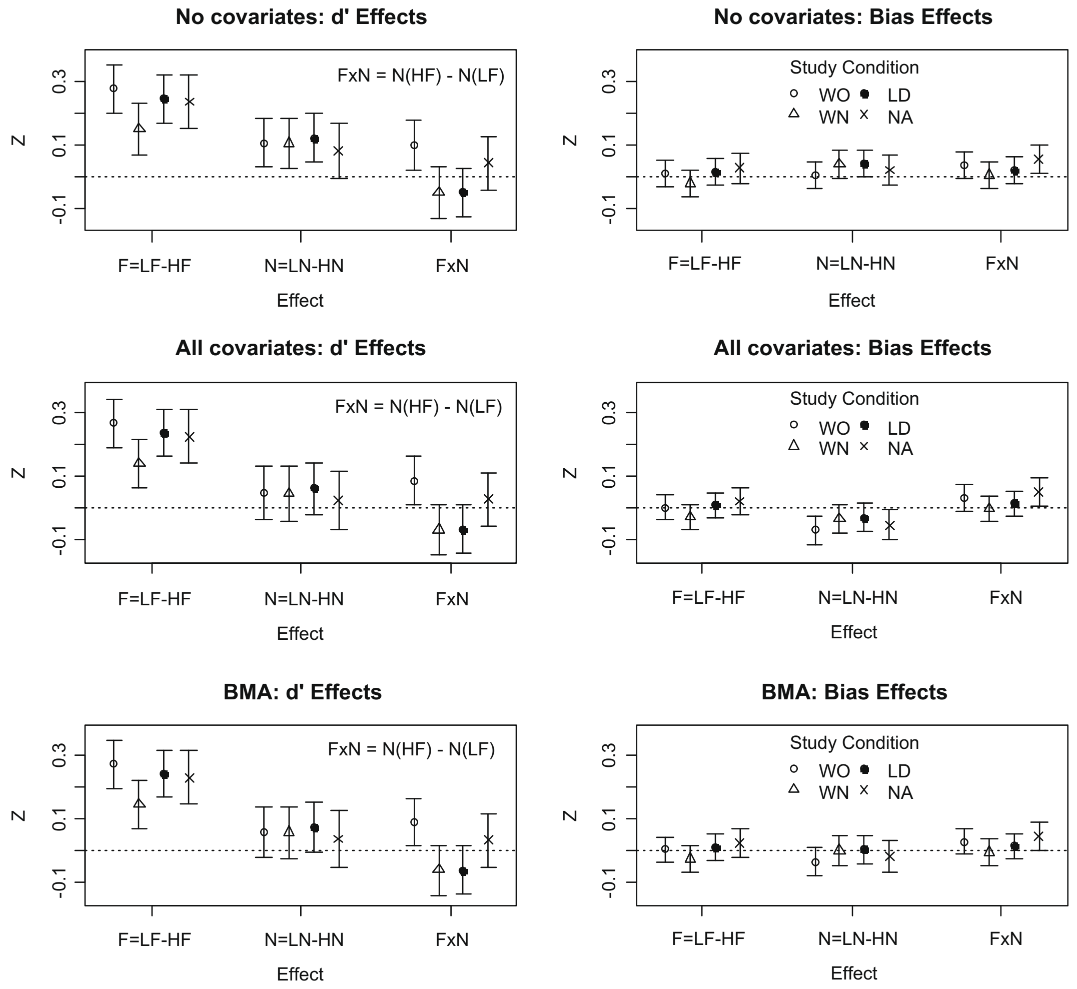

Fig. 4. Fixed effect estimates for episodic memory with 95% confidence intervals. Reliable effects (at the $p < .05$ level) are indicated by intervals that do not cross the dotted line indicating a zero effect. Item conditions: F = word frequency, N = neighborhood density, L = low and H = high. Study conditions: WO = words-only, WN = word–nonword, LD = lexical decision and NA = naming. BMA = Bayesian Model Average.

finding no density effects in lexical decision, suggesting that episodic density effects may be found here.

Recognition memory performance

As shown in Fig. 2, overall episodic accuracy was good ($d' = 1.6$) and there was a small but reliable bias of about 5% towards new responses ($c = 0.12$, reliably than greater zero, $p < .001$). Relative to the word–nonword condition, d' was reliably greater for naming (by 0.2, $p = .013$), marginally less for words-only (by 0.14, $p = .052$) and slightly but not reliably less for lexical-decision (by 0.06, $p = .47$). Relative to the word–nonword condition, new bias was reliably greater for words-only (by 0.11, $p = .028$), marginally less for lexical-decision (by 0.095, $p = .058$) and slightly but not reliably less in the naming condition (by 0.03, $p = .57$).

Bias was largely equated across item set conditions. Only two bias effects related to item sets were reliable, a greater new bias for low than high density (by 0.04) for lexical-decision study ($p = .047$) and an interaction in naming where new bias increased with density for low frequency words but decreased with density for high frequency words ($p = .022$, difference = 0.05). Hence, the results for bias are largely as predicted by likelihood theories.

Although word frequency and density effects generally followed a mirror pattern, with low frequency/density words having more hits and fewer false alarms than high density/frequency words, there were exceptions. The frequency mirror effect most often failed for new items, particularly for low density words in the words-only and naming conditions and for both high and low density words in the lexical-decision condition. The density mirror effect also failed for new low frequency items in the words-only and naming conditions and for old high frequency items in all but the words-only condition.

Covariate effects

We fit all covariate models and calculated p_{BIC} of each within the set of 128. No-covariate model was associated with positive or stronger evidence, consistent with some model uncertainty, although models containing only inter-

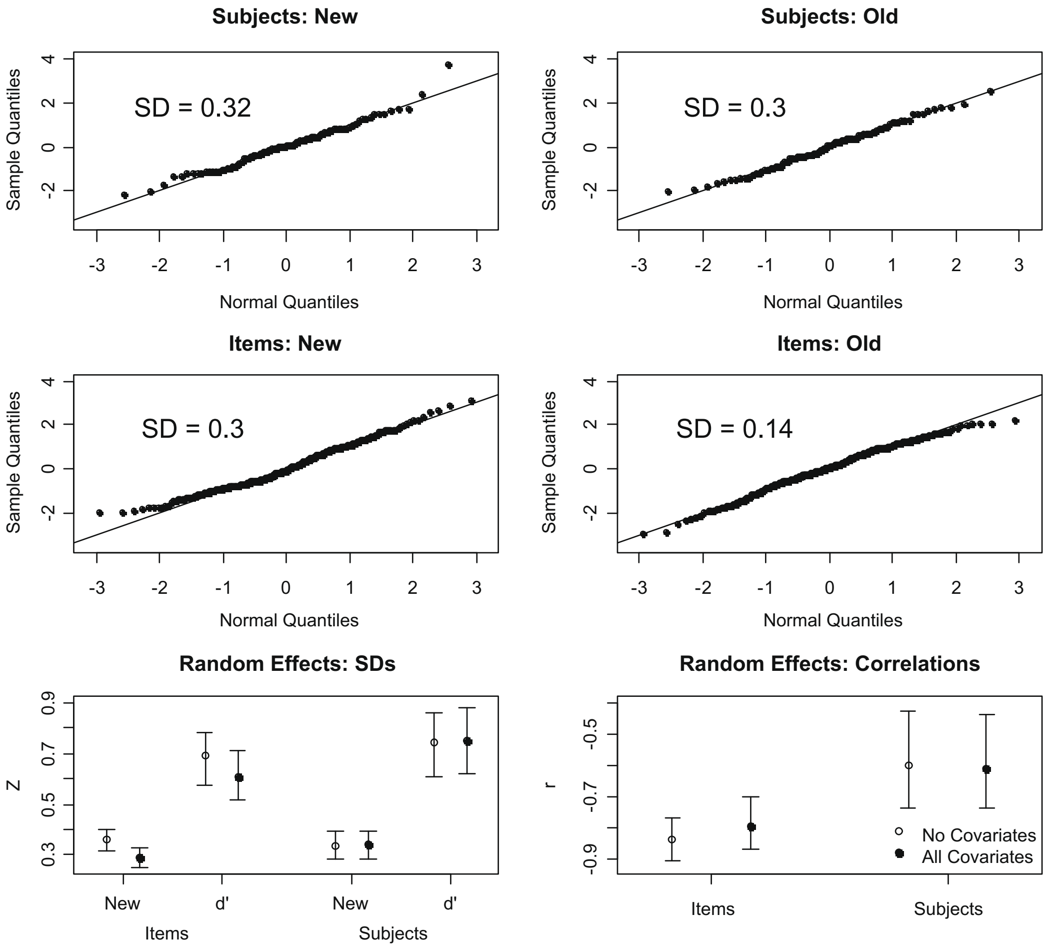

Fig. 5. The upper two rows show sample random effects as normal quantile–quantile plots of best linear unbiased predictors (BLUPs) for individual subjects and items broken down by test-type (new or old items), and their SDs, for the model with no covariates. The bottom row shows population *SD* and correlation estimates for the new condition and for *d*′, and associated 95% confidence intervals obtained from fits to 500 bootstrap samples from the no covariate and all-covariates models (where these effects are reliable at the $p < .05$ level intervals do not include zero).

item similarity covariates were most likely. In particular, the model including orthographic stem and Levenshtein covariates was most favoured ($p_{\text{BIC}} = .70$), followed by the model with an orthographic stem similarity covariate alone ($p_{\text{BIC}} = .16$) or in combination with phonological stem similarity ($p_{\text{BIC}} = .07$) or all three ($p_{\text{BIC}} = .05$). No other model had $p_{\text{BIC}} > .01$, including the no-covariate model, indicating that the inclusion of at least some covariates is strongly supported (i.e., $p_{\text{BIC}} > .99$ for all covariate models vs. the no-covariate model).

The top row of Fig. 3 shows the estimated effects of each covariate assuming no-covariate model uncertainty (i.e., assuming that the covariate does enter the model). They show that, even if letter frequency and neighborhood overlap are assumed to be in the model, they do not have effects that reliably differ from zero. For bigram frequency only the bias effect is reliable, whereas all of the remaining covariates have reliable effects on both *d*′ and bias.

However, when model uncertainty is fully taken into account, only orthographic stem and Levenshtein covariates were supported, as shown in the bottom left of Fig. 3. In particular $p_{\text{BIC}} > .99$ for the stem metric, indicating very strong evidence, and $p_{\text{BIC}} = .76$ for the Levenshtein metric, indicating positive evidence. The effect estimates for these variables indicate that *d*′ decreases by 0.07 and 0.03 for each one *SD* increase in item noise as measured by the stem and Levenshtein metrics, respectively (see Table 1 for *SD* values). These estimates are in the direction predicted by item-noise theories, a decrease in *d*′ as item noise, as measured by inter-item similarity, increases. Orthographic stem and Levenshtein metrics had the opposite effect on bias, an increase by 0.10 and 0.08 per *SD*, respectively.

Word frequency and density effects

The top row of Fig. 4 shows the effect estimates and tests for the no-covariate model, corresponding to the results shown in Fig. 2. For *d*′, word frequency main effects were reliable in each study condition (all $p < .001$) with low greater than high frequency by 0.22 overall ($p < .001$). The density main effects were reliable for all but naming, where it was marginal ($p = .065$, $p < .01$ otherwise), with an average magnitude of 0.1 ($p < .001$). For words-only and naming, Fig. 2 shows an interaction between density and frequency caused by a lack of density effects for false alarms and d'. The interaction in d' was reliable for words-only ($p = .012$), as were interaction contrasts between the word–nonword study condition and the words-only ($p = .002$) and lexical-decision ($p = .041$) study conditions.

Fig. 4 also shows findings for the model with all covariates included. The covariates greatly attenuated the effect of density on accuracy, so the average main effect in d' across study conditions was only 0.04 ($p = .16$). No d' main effects of density within study conditions were reliable. In contrast, the frequency main effects remained unchanged in magnitude and highly reliable in each study condition. However, the interaction between density and frequency in the words-only condition ($p = .028$) and the contrasts between interactions in the word–nonword study condition and the words-only ($p = .003$) and lexical-decision ($p = .038$) study conditions remained reliable. Consistent with the negligible association between covariates and frequency effects, the exceptions to the frequency mirror pattern shown in Fig. 1 were also evident in plots (not shown) of hit and false alarm rates adjusted for covariate effects.

The bottom row of Fig. 4 displays BMA estimates of word frequency and density effects. The same conclusions about these effects continue to hold when covariate model uncertainty is taken into account: for d', word frequency had strong main effects, density did not have any reliable main effects, but it did reliably interact with frequency in the words-only condition. Bias effects were also similar, except that the main effect of density in the words-only condition, and the interaction between density and frequency in the naming condition were no longer reliable. The same conclusions held for all of the more probable ($p_{\mathrm{BIC}} \geqslant .05$) individual covariate models, except when only stem similarity was a covariate. In this case, density main effects remained reliable for the words-only ($p = .048$) and lexical-decision ($p = .017$) conditions and marginal for the words–nonword condition ($p = .064$).

Random effects

BIC selected a model with subject and item random effects that differed as a function of test-types, and which also allowed for correlations between random test-type random effects. In particular, this model was preferred ($p_{\mathrm{BIC}} > .99$) relative to the five nested models that omitted some or all of these random effect parameters. We were unable to find any other factor, or combination of factors, that improved BIC relative to this model. In particular, BIC did not support different random effect parameters for either word frequency or density, suggesting that previously reported set-level effects of these variables were not due to differences in item variability. However, in agreement with Rouder and Lu (2005), omitting item random effects resulted in an underestimate of accuracy as measured by an overall $d' = 1.5$ estimate for the model with only a single random subject effect compared to an overall $d' = 1.6$ estimate for the selected model.

Fig. 5 also shows the best linear unbiased predictors (BLUPs, Baayen et al., 2008) of the random effects for each item and subject as normal quantile–quantile (QQ) plots. The QQ plots enable a check of the assumption that these random effects are normally distributed as well as a check on whether the data set contained any influential outliers. Consistent with the assumption that the random effects are normally distributed, the BLUPs lie along the main diagonal of the normal quantile–quantile plots in Fig. 5. Fig. 5 also shows that our results are not distorted by a few influential outlying items or subjects. We also found that analyses that did not make the assumption of normally distributed item and subject effects (i.e. analyses that calculated estimates for items by aggregating over subjects and estimates for subjects by aggregating over items) were consistent with conclusions based on the BLUPs.

The bottom row of Fig. 5 shows population estimates of random effect standard deviation estimates for the effect of study (i.e., d', the difference between new and old conditions caused by study) and the new condition, and their correlation. We report results for this parameterization, rather than the bias and d' parameterization used to report fixed effects, as it is directly interpretable in terms of subject and item effects without study and the effect of study alone (i.e., $d' = \text{old–new}$). In terms of our earlier illustration of the negative correlation predicted by likelihood theories, where subjects or items A and B have new estimates of 1 and 2 and old estimates of 4 and 3, respectively (i.e., a mirror effect), corresponding study effect estimates are 3 and 1. Hence, new and study estimates are also predicted to be negatively correlated.

Estimates of new and d' standard deviations, and correlations, are displayed side-by-side for the models with no covariates and all covariates. As would be expected, the inclusion of item covariates reliably reduced item variance and the correlation, but does not effect subject estimates. Because the presence or absence of covariates did not otherwise affect the pattern of random effect results we discuss both together.

Study effects (d') were considerably more variable than new effects for both subjects and items, but this did not result in greater variability in the old condition because of the strong negative correlations between new and d' variability. These correlations indicate that study had a greater effect for subjects and items with lower values when new, as predicted by likelihood theories. For subjects, the correlation was sufficiently strong to cancel the variability in the old condition caused by the addition of d', resulting in approximately equal standard deviations in new and old conditions (see BLUP plots for corresponding sample estimates). For items, the correlation is even stronger, resulting in old item standard deviations that are less than half new item standard deviations.

Summary

Our results produced clear answers to the six sets of questions which we hoped to answer. With regard to the questions related to random effects the following answers were found. Both subject and item variability reliably affected recognition memory performance, with subject variation larger than item variation. These results indicate that analyses with both types of random effects are preferable to usual set-level analyses. There was also clear evidence that both random item and subject effects were a function of test-type, with the affect of study being negatively correlated with the level before study (i.e., in the new condition), consistent with the predictions of likelihood theories. However, item variation cannot explain previous set-level zROC slope results. If anything, confounding by item variation in set-level analyses leads to an underestimate of the degree to which old variation is greater than new variation, due to strong negative correlations. Word frequency and density effects did not enter into the random effects model, so effects of these variables on set-level analyses cannot be attributed to item variability differences.

With regard to fixed effects the following answers were found. First, the word frequency and density effects previously found with set-level analysis are reliable when random item effects are taken into account, so neither are due to Type 1 errors resulting from the overestimation of reliability by set-level analysis. Episodic recognition was not found to be influenced by the frequency of word constituents (letters and bigrams) but was clearly influenced by inter-item similarity covariates. The direction of the latter effect was as predicted by item-noise theories, a decrease in accuracy as inter-item similarity, and hence item noise, increases. These conclusions hold even when a proper account is taken not only of both item and subject random effects but also of covariate model uncertainty. Finally, only word frequency continued to have an effect when performance was adjusted for differences between item sets in inter-item similarity covariates. These results, along with the finding that density did not have a reliable effect on performance in lexical tasks, clearly support an item noise explanation of density effects. In contrast, word frequency effects were unaffected by the inclusion of inter-item similarity covariates. We now discuss the implications of all of these findings.

Discussion

Our finding of strong inter-item similarity effects expands to words Cleary et al.'s (2007) findings about inter-item similarity effects for nonwords. As well as validating an orthographic version of Cleary et al.'s stem inter-item similarity measure, we also found reliable effects of other measures of inter-item similarity, particularly orthographic Levenshtein similarity. These results extend Yarkoni et al.'s (2008) findings that Levenshtein distance provides a useful similarity metric for lexical decision and naming to the domain of recognition memory for words.

In light of these results, further research on the influence of similarity on other metrics seems warranted, such as semantic similarity, which is likely to play an important role in recognition memory (e.g., Roediger & McDermott, 1995). Although phonological similarity measures did not have a strong effect in our paradigm, they may warrant further investigation in other paradigms (e.g., auditory presentation of test words). We also note that despite neighbor overlap not playing a large role in multiple-covariate models, it did produce the greatest attenuation of density main effects on d' (0.03 overall) of any single covariate model.

The strong orthographic inter-item similarity effects that we observed are consistent with recognition memory being affected by item noise (Criss & Shiffrin, 2004) and with Malmberg and Nelson's (2003) emphasis on the importance of word identification processes when study duration is relatively short (see also Criss & Malmberg, 2008). On the empirical front, our results suggest that previously reported effects of density (Cortese et al., 2004, 2006; Glanc & Greene, 2007; Heathcote et al., 2006) may be to a large degree the result of item noise, at least when it is quantified by both stem and Levenshtein measures. However, further research is required to confirm this suggestion, particularly for the Cortese et al. studies where item noise was minimized using a stimulus set in which no two items shared both the same orthographic and phonological rime, and only two pairs of items contained the same phonological rime.

On the theoretical front, the item noise effects observed in the present experiments might be accommodated within the BCDMEM context noise model (Dennis & Humphreys, 2001) if a sufficiently large set of word nodes were partially activated during the identification process, as suggested by Lamberts et al.'s (2003) feature-sampling model of recognition memory, and a variety of reading models (Andrews, 1997). Partial activation could result in stronger associations to the study context for items with higher inter-item similarity, and hence a reduction in accuracy relative to items with lower inter-item similarity.

The strong inter-item similarity effects that we observed reinforce Cleary et al.'s (2007) methodological caution about potential confounding in the interpretation of item set manipulations due to inter-item similarity differences. For example, analysis of the items used by Malmberg et al. (2002) show that their letter frequency manipulation was strongly confounded with inter-item orthographic stem similarity. Out of a possible 71 matches, low frequency words matched on average 0.85 and 2.4 times for low and high letter frequency, respectively, and for high frequency words 0.53 and 2.33 times. In this case, the interpretation would remain consistent with a general item-noise account, but in other cases the interpretation of item set manipulations might be substantially altered (see Cleary et al., 2007). Although we did not find any reliable effects of letter frequency using items with an almost identical mean frequency to Malmberg et al., their high-low item set manipulation spanned a much larger range of letter frequencies than our item set. Hence, our null finding may have been the result of the limited range over which letter frequency varied in our items.

At a more general level, the success of our covariate analyses encourages the use of item covariates, rather than the traditional high-low item-set design, to investigate item effects on recognition memory. Baayen, Levelt, Schreuder, and Ernestus (2008) came to a similar conclusion in relation to lexical memory, based on a re-analysis of data from a high-low item-set design reported by Baayen, Dijkstra, and Schreuder (1997). Dichotomization discards information in the covariate measure and so may account for less variance in the data. To investigate whether this was the case in our episodic memory data we replaced item-set (word frequency and density) terms in the all-covariates model with word frequency, or its logarithm, as a covariate. BIC decreased substantially, with the logarithmic covariate having the lowest BIC. The logarithmic word frequency covariate had an estimated effect ($d' = 0.12$) slightly larger than the best other covariate, stem orthographic similarity. These results suggest that power could be improved in episodic recognition experiments if log-frequency, as well as inter-item similarity measures, were routinely used as covariates, even in experiments not focused on testing item effects.

Perhaps even more important, we note that Levenshtein similarity was reliably associated with the effect of density, even though our low and high density item sets were virtually equivalent on mean Levenshtein similarity (see Table 1). This strongly suggests that the traditional approach of controlling item confounds by equating mean values between item sets is inadequate, and that instead a covariate approach is preferable. In future work this sort of correlation analysis could be employed to investigate whether the effects of the inter-item similarity variables observed here can be explained, for example, by alternative measures of the similarity of words to all other words in the lexicon other than the density and frequency measures used here.

Our findings that word frequency strongly affected RT in naming and lexical-decision tasks are consistent with an explanation of the word frequency effect in episodic recognition in terms of study attention differences (Glanzer & Adams, 1990; Malmberg & Nelson, 2003). The fact that the word frequency effect was not attenuated by our inter-item similarity covariates, which are a measure of item noise, is consistent with word frequency being a context noise effect (Dennis & Humphreys, 2001) or due to differences in recollection (Joordens & Hockley, 2000) or study attention. In contrast to word frequency, density did not reliably affect naming and lexical-decision RT. As these results indicate no difference in identification processes for low and high density words, the attention hypothesis predicts no density effect in recognition memory. This prediction was confirmed when orthographic inter-item similarity was controlled, which is consistent with density effects in episodic recognition being due to item noise caused by orthographic inter-item similarity.

It is possible that at least some of the frequency effect may be caused by a type of item noise not controlled by our covariates, such as semantic similarity or semantic feature frequency, although Criss and Malmberg (2008) provide evidence that semantics do not mediate word frequency effects in hit rates. From the point of view of item-noise theories, our inter-item similarity metrics are relatively crude, as they measure the average similarity of an item to all other experimental items, whereas these theories predict effects of similarity to only previously studied items. Such metrics will be correlated with the effects of similarity to previously studied items at the population level, due to random allocation of items to conditions. However, item-noise models should predict stronger effects for refined metrics that take a weighted (perhaps study-test lag dependent) sum of similarities over previously studied items. In future research, the item covariate approach we developed here could easily be adapted to make such a direct test, and to elaborate previously difficult-to-explore theoretical details, such as the form of the weighting function.

Our results are consistent with the idea that recognition decisions are based on likelihood or some other sophisticated decision process which takes account of the mnemonic characteristics of subjects and items. The fact that we observed this pattern not only at the level of classes of items but also for individual items (i.e., a strong negative correlation between memory evidence for new items and the effect of study on memory evidence) is consistent with a strong version of the likelihood decision model that takes account of individual item and subject characteristics. In any case, the result is a well calibrated decision in which bias is equated across different types of items.

In contrast to likelihood theories, two process theories of the mirror effect (e.g., Joordens & Hockley, 2000; Reder et al., 2002) would require a serendipitous balance of their effects to explain equal bias across different item types. This suggests that the two process model may benefit from assuming an appropriate dependence between processes. A more radical elaboration, furthering the evolution of dual-process (familiarity and recollection) theory suggested by Wixted (2007), would be to assume that subjects base their recognition decisions on a likelihood estimate that is informed by knowledge about item and subject characteristics, both in terms of familiarity and the probability of recollection.

Although our bias and random effects results are consistent with likelihood theories, our failure to find a full frequency mirror effect in some cases may bring this conclusion into question. In most cases, failure of the frequency mirror effect is restricted to hit rates (McClelland & Chappell, 1998), whereas we mainly observed failures of the false alarm portion of the frequency mirror effect. Dual-process theory can explain the failures for hit rates as due to the effect of recollection not being sufficient to overcome the effect of familiarity. However, it cannot explain the failures for new items unless familiarity is affected by the study task, which is inconsistent with the general attribution of familiarity to pre-experimental factors. Likelihood theories can accommodate failures of the mirror effect due to differences in bias (Glanzer, Hilford, & Maloney, 2009), but as our estimates of bias did not differ between item sets, an approximate rather than exact likelihood calculation is implicated.

An approximate calculation is consistent with “subjective” likelihood theories (Criss, 2006; Criss & McClelland, 2006), which assume that the subject is not fully informed about all of the relevant properties of a test item. In particular, models such as REM, SLiM and BCDMEM assume that subjects do not have exact knowledge about variations in the effect of study (but see also Glanzer, Adams, Iverson, & Kim, 1993, for a discussion of the effects of such inaccurate knowledge in ALT). Hence, item specific variations, such as the study attention differences suggested by Malmberg and Nelson (2003), may introduce inaccuracies into the likelihood calculation, causing the complete mirror effect pattern to be violated. Clearly, however, these possibilities are speculative and a fuller evaluation will require fits of these specific models to data.

The fact that we found old item variance to be *less* than new item variance is inconsistent with ROC analyses that aggregate over items. Such analyses have consistently found zROC slopes around 0.8, indicating *greater* old than new variance. Ratcliff et al. (1992) suggested that "there might be a large enough difference among individual study or test items so that mixing them would produce slopes around 0.8″ (p. 527). Our findings indicate that item variability is actually much less for old than new items, and so it cannot explain the ROC results. However, concerns have been expressed about bias and variability in slope estimates (MacMillian, Rotello, & Miller, 2004), and both Mueller and Weidemann (2008), Ratcliff and Starns (2009) have questioned whether slopes accurately reflect variance in memory evidence.

If study does induce greater variance, as suggested by results from set-level ROCs, our analyses will underestimate *d*′, and bias estimates will also be affected (e.g., Grider & Malmberg, 2008). Equally, however, our results indicate that estimates of these quantities based on a zROC that aggregates over items will be in error, because zROC slopes will be biased upward by lower old than new item variance. Clearly more research is needed to resolve these issues. One possibility that extends the approach we have taken here is to model systematic effects on variability using covariates specific to old items, such as study-test lag or measures of attention during study. Such covariates would be expected to cause increased variability in memory strength for old items relative to new items. Another possibility is to develop random item and subject effect analyses suitable for ROC data. Although such analyses are challenging, as they require nonlinear models, Morey, Pratte, and Rouder (2008) present a promising Bayesian approach to their estimation. Ideally both approaches could be combined, allowing item covariates to be applied to ROC analyses.

In conclusion, our results strongly support the existence of large and reliable item effects in recognition memory (Lamberts et al., 2003; Rouder & Lu, 2005). In particular, we found clear support for Rouder et al.'s (2007) correlated random effects model. We have also extended their approach by showing that item covariates account for a substantial amount of variance in recognition memory performance. Hence, there appears to be a strong case for the routine use of mixed models with item covariates for the analysis of recognition memory data. This approach not only has the potential to increase statistical power, but also to allow for control of confounding item characteristics and to provide theoretical insights into sources of noise and the decision processes that govern recognition memory. In particular, it allowed us to show that the density effect is the result of item noise as measured by orthographic inter-item similarity, that the word frequency effect is not explained by orthographic and phonological inter-item similarity, and that subjects are able to take account of their mnemonic ability, and the mnemonic characteristic of items, as predicted by likelihood theories.

Acknowledgments

Thanks to Harald Baayen, Michael Cortese and four anonymous reviewers, and to the Macquarie Centre for Cognitive Science for making the ARC Nonword Database available on the web (http://www.maccs.mq.edu.au/~nwdb/), to Sally Andrews for advice on item selection and Jeff Rouder for comments on statistical issues. We acknowledge support from an ARC Discovery Grant (DP0451223).

Appendix A. Item sets

A.1. Low frequency, low neighborhood density words

Afar, blur, chef, defy, dual, glee, grub, hazy, isle, itch, jolt, knit, levy, loaf, menu, mesh, numb, oval, oxen, putt, soak, turf, void, wolf, wrap, abide, baton, clues, cured, derby, disks, easel, farce, hosts, hymns, idiom, irons, joins, knack, lawns, lions, marsh, mirth, ounce, plush, quart, reins, roast, sofas, truce, bolted, chewed, clique, compel, darted, draped, flexed, fluent, glazed, glowed, hinted, inform, larvae, masked, melted, nudged, pearls, relied, risked, spiced, summed, taxing, undone, wedged, weighs.

A.2. Low frequency, high neighborhood density words

Cove, flaw, gilt, hack, hike, leak, melt, mule, rack, reed, swan, tilt, vent, wail, wink, worm, yell, zeal, mend, coil, lure, curl, teen, ramp, hush, berry, bully, forts, gears, grate, hatch, latch, liner, merry, minus, mouse, noses, packs, pears, pines, rover, rowed, scare, shave, stale, stark, tunes, vines, waded, wakes, banker, cocked, dating, fender, hailed, heaped, jagged, jailed, lashes, licked, mashed, paving, potted, ragged, raving, ringed, rocket, shakes, singer, soared, spared, tipped, wailed, washes, winged.

A.3. High frequency, low neighborhood density words

Acts, bird, club, copy, desk, drew, drop, duty, edge, eggs, ends, evil, film, glad, golf, iron, join, knee, onto, plus, self, sign, term, tree, unit, ahead, block, brief, noted, facts, chief, claim, dance, enemy, yards, tools, ideal, image, limit, loose, march, liked, funds, queen, names, towns, serve, steel, thick, doors, agreed, belong, cattle, caused, dreams, effort, engine, gained, handed, helped, latest, lifted, motion, nodded, normal, proved, raised, remain, rolled, smiled, strain, talked, tongue, weight, wished.

A.4. High frequency, high neighborhood density words

Blow, boat, coal, coat, ease, feed, fill, king, list, mark, milk, nose, note, park, pool, post, push, race, rise, safe, ship, sold, test, tour, yard, beach, bound, faced, forms, games, grown, hills, homes, lines, loved, lower, match, moves, pages, paint, parts, rates, river, sales, sight, store, track, train, walls, waves, bitter, failed, faster, filled, formed, latter, leaned, locked, master, missed, packed, paying, picked, pulled, shared, showed, stared, stated, stayed, waited, washed, leader, rights, finger, waters.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jml.2009.09.004.

References

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*, 439–461.

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 514–544.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Baayen, R. H., Levelt, W. M. J., Schreuder, R., & Ernestus, M. (2008). Paradigmatic structure in speech production. *Proceedings of the Chicago Linguistics Society, 43*, 1–28.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language, 36*, 94–117.

Bates, D. M. (2005). Fitting linear mixed models in R. *R News, 5*, 27–30.

Buchta, C. & Hahsler, M. (2007). cba: Clustering for business analytics. R package version 0.2-4.

Clark, H. H. (1973). The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Cleary, A. M., Morris, A. L., & Langley, M. M. (2007). Recognition memory for novel stimuli: The structural regularity hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 379–393.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). New York: Academic Press.

Cortese, M. J., Watson, J. M., Wang, J., & Fugett, A. (2004). Relating distinctive orthographic and phonological processes to episodic memory performance. *Memory & Cognition, 32*, 632–639.

Cortese, M. J., Watson, J. M., Khanna, M. M., & McCallion, M. (2006). Revisiting distinctive processes in memory. *Psychonomic Bulletin & Review, 13*, 446–451.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*, 461–478.

Criss, A. H., & Malmberg, K. J. (2008). Evidence in favor of the early-phase elevated-attention hypothesis: The effects of letter frequency and object frequency. *Journal of Memory and Language, 59*, 331–345.

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55*, 447–460.

Criss, A. H., & Shiffrin, R. M. (2004). 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*, 778–786.

Davis, C. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behaviour Research Methods, 37*(1), 65–70.

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

Estes, W. K., & Maddox, W. T. (2002). On the processes underlying stimulus-familiarity effects in recognition of words and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 1003–1018.

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

Glanc, G. A., & Greene, R. L. (2007). Orthographic neighborhood size effects in recognition memory. *Memory & Cognition, 35*, 365–371.

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

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 5–16.

Glanzer, M., Adams, J. K., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 81–93.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*, 546–567.

Glanzer, M., Hilford, A., & Maloney, L. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review, 16*, 431–455.

Grider, R. C., & Malmberg, K. J. (2008). Discriminating between changes in bias and changes in accuracy for recognition memory of emotional stimuli. *Memory & Cognition, 36*, 933–946.

Heathcote, A., Ditton, E., & Mitchell, K. (2006). Word-frequency and word-likeness mirror effects in episodic recognition memory. *Memory & Cognition, 34*, 826–838.

Joordens, S., & Hockely, W. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1534–1555.

Lamberts, K., Brockdorff, N., & Heit, E. (2003). Feature-sampling and random-walk models of individual-stimulus recognition. *Journal of Experimental Psychology: General, 132*, 351–378.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behaviour, 12*, 119–131.

Malmberg, K. J., & Nelson, T. O. (2003). The word-frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition, 31*, 35–43.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30*, 607–613.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A users guide* (2nd ed.). New York: Cambridge University Press.

MacMillian, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychophysics, 66*, 406–421.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 734–760.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology, 52*, 376–388.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*, 465–494.

Myung, I.-J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79–95.

Quene, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*, 413–425.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org>.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 Nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology A, 55*, 1339–1362.

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

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*, 59–83.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. A. (2002). A reexamination of stimulus-frequency effects in recognition: Two mirrors for low- and high-frequency pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 138–152.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

Rouder, J., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*, 573–604.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process-dissociation model. *Journal of Experimental Psychology: General, 137*, 370–389.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika, 72*, 621–642.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92*, 351–370.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.

Wagenmakers, E.-J., & Brown, S. D. (2007). On the linear relationship between the mean and the standard deviation of a response time distribution. *Psychological Review, 114*, 830–841.

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

Yarkoni, T., Balota, D., & Yap, M. (2008). Behavioral and neural explorations of a new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971–979.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.