

Word frequency and word likeness mirror effects in episodic recognition memory

ANDREW HEATHCOTE, ELIZABETH DITTON, and KRISTIE MITCHELL
University of Newcastle, Callaghan, Australia

Estes and Maddox (2002) suggested that the word frequency mirror effect in episodic recognition memory might be due to word likeness rather than to the frequency of experience with a word per se. We examined their suggestion using a factorial manipulation of frequency and neighborhood density, a measure used in lexical memory research to measure orthographic word likeness. For study with no specified task, main effects of density and frequency were in the mirror order, confirming the hypothesized mirror effect of word likeness but not its role in producing the frequency mirror effect. Lexical decision study increased the size of both mirror effects, even though the density manipulation had a negligible effect on lexical decision performance for words. Post hoc analyses showed that neither mirror effect could be explained by differences in lower order measures of word likeness (letter and bigram frequency). The joint orders of frequency and density results were mirrored across new and old conditions in accordance with attention likelihood theory (ALT), but density effects on z -ROC slope suggest that ALT may require extension to accommodate the effect of word likeness on response confidence.

Lexical memory contains the orthographic, phonological, and semantic representations used, for example, during reading and speaking. Episodic memory, in contrast, is required to decide whether or not an item, such as a word, was experienced in a particular context, such as the last study list. The lexical characteristics of words can have a marked effect on episodic memory. The most studied example is *normative word frequency*, or the number of times a word occurs in a text corpus (see, e.g., Kučera & Francis, 1967). In the experiments reported here, we examine the origins of the word frequency effect in episodic recognition memory and its relationship to the word frequency effect in lexical memory.

Word frequency has robust and opposite effects in cued lexical and episodic memory tasks. High-frequency words produce faster and more accurate naming and lexical (word vs. nonword) decisions than low-frequency words (Andrews, 1992; Balota & Chumbley, 1984; Schilling, Rayner, & Chumbley, 1998). In contrast, episodic recognition accuracy is better for low- than for high-frequency words. Decreased episodic accuracy commonly results from de-

creased performance for both unstudied (new) and studied (old) words. This pattern of results is called a *mirror effect* because higher false alarm rates (FARs) are mirrored by lower hit rates (HRs) for high-frequency in comparison with low-frequency words. The frequency mirror effect has been found by many researchers using a variety of item sets and recognition paradigms (Glanzer & Adams, 1985), leading Glanzer, Adams, Iverson, and Kim (1993) to describe it as one of the regularities of recognition memory.

Differences between high-frequency and low-frequency words postulated to cause the mirror effect are almost as numerous as the empirical findings. In comparison with low-frequency words, high-frequency words have been hypothesized to (1) have more lexical features (Glanzer & Bowles, 1976; McClelland & Chappell, 1998) because they have more dictionary definitions (see, e.g., Reder, Anderson, & Bjork, 1974) and produce more distinct associative responses (see, e.g., Paivio, Yuille, & Madigan, 1968); (2) be associated to more episodic contexts (Dennis & Humphreys, 2001; Reder et al., 2000; Sikström, 2001); (3) be composed of less distinctive lexical features (Shiffrin & Steyvers, 1997); (4) have greater baseline (prestudy) memory strength in either lexical memory (Murdock, 2003; Reder et al., 2000) or episodic memory (Glanzer & Bowles, 1976; Wixted, 1992); (5) have a slower episodic learning rate (Glanzer & Adams, 1990; Glanzer & Bowles, 1976; McClelland & Chappell, 1998; Murdock, 2003); (6) be more difficult to recollect (see, e.g., Joordens & Hockley, 2000; Reder et al., 2000; Yonelinas, 1994); and (7) have a greater degree of word likeness (Estes and Maddox, 2002). In this article, we investigate the last hypothesis.

Estes and Maddox (2002) defined word likeness as “an index of the degree to which an item is typical of the broad

Thanks to Kerry Chalmers, Bill Hockley, Emily Bohlscheid, two anonymous reviewers, and Ben Murdock for helpful comments. Thanks also to the Department of Psychology, University of Western Australia, for making the MRC Psycholinguistic Database (www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm) available on the Web, and to the Macquarie Centre for Cognitive Science for making the ARC Nonword Database (www.maccs.mq.edu.au/~nwdb/) similarly available. Correspondence concerning this article should be addressed to A. Heathcote, School of Behavioural Sciences, Aviation Building, University of Newcastle, Callaghan, NSW 2308, Australia (e-mail: andrew.heathcote@newcastle.edu.au).

Note—This article was accepted by the previous editorial team, when Colin M. MacLeod was Editor.

superordinate category word” (p. 1014). They conjectured that the frequency effect was mediated by word likeness, and specifically orthographic word likeness, rather than being a direct effect of the frequency of experience with a particular word, on the basis of their finding of a parallel rather than a mirror effect of different levels of familiarization training on later episodic recognition memory performance (see also Chalmers & Humphreys, 1998; Maddox & Estes, 1997). A parallel effect occurs when FAR and HR effects are in the same direction rather than opposite directions. Estes and Maddox found that words experienced more frequently during familiarization training (the experimental analogue of high natural language frequency) had both higher HRs and higher FARs than words experienced less frequently during familiarization.

In agreement with Estes and Maddox’s (2002) conjecture, Zechmeister (1972) found that words rated as highly distinctive produced better accuracy and a mirror effect relative to less distinctive words. Malmberg, Steyvers, Stephens, and Shiffrin (2002) found that average letter frequency, a low-order measure of orthographic word likeness, produced a mirror effect, with better accuracy for words containing less frequent letters. Malmberg et al. also found that when letter frequency differences were controlled, the word frequency mirror effect remained, rather than yielding to a parallel effect, as might be predicted on the basis of Estes and Maddox’s explanation if letter frequency alone were sufficient to account for word likeness. However, average letter frequency is a measure of word likeness at the level of a word’s smallest parts, so it remains possible that a parallel frequency effect could emerge if properties of higher order configurations of letters were controlled.

In our experiments, neighborhood density (Coltheart, Davelaar, Jonasson, & Besner, 1977) was used as a higher order measure of orthographic word likeness. The neighborhood density of a given word is the number of words that can be made by changing one letter of that word (e.g., *cat* has the neighbors *sat*, *cut*, *cap*, etc.). Hence, for a word of length L (e.g., $L = 3$ for *cat*), density is a measure of the position-specific frequency of that word’s highest order substrings, of length $L - 1$ (i.e., *_at*, *c_t*, *ca_*), in words of the same length. In our first experiment, we tested the

hypothesis that frequency effects are caused by word likeness by factorially manipulating frequency and density. If Estes and Maddox’s (2002) hypothesis is correct, controlling for density may remove the frequency mirror effect and perhaps produce a parallel effect in line with their familiarization results. A secondary aim was to determine whether or not density, like the lower order letter frequency measure of word likeness, produces a mirror effect.

OVERVIEW OF THE EXPERIMENTS

In both experiments reported here, a 2×2 design was used, with high-frequency and low-frequency words crossed with high-neighborhood-density and low-neighborhood-density words (see Table 1 and the Appendix). In this design, the main effects of frequency and density test the effects of each factor with the other controlled. In both experiments, participants provided ratings of confidence in their recognition memory judgments. We examined both mean confidence judgments (R) for new and old items and the probability of judging a test item as old [$p(\text{Old})$] for new items (i.e., FARs) and for old items (i.e., HRs) to determine whether our frequency and density manipulations produced parallel effects or mirror effects.

Word frequency was measured using the Kučera and Francis (1967) norms obtained from the MRC psycholinguistic database, Version 2.00 (Coltheart, 1981). Neighborhood density was determined using the Neighs program.¹ Average values are reported in Table 1. The density counts are based on a relatively small corpus, on the order of 10,000 words, that avoids rare words unlikely to be in participants’ vocabularies, but more words could arguably be included. For example, the N-Watch program (Davis, 2005) uses approximately three times as many words and also reports word frequency on the basis of the more extensive CELEX corpus (Baayen, Piepenbrock, & van Rijn, 1993). However, average density and frequency for our word sets based on N-Watch did not differ much from those reported in Table 1, except that, as might be expected, density values were slightly higher overall.

The stimuli used in both experiments were 560 words, of which half were high frequency and half were low frequency. For both high- and low-frequency word sets, there

Table 1
Mean Measures for Each Word Type

Measure	Low Frequency				High Frequency			
	LN	% of Words	HN	% of Words	LN	% of Words	HN	% of Words
Density	1.36	100	4.75	100	1.49	100	4.74	100
ln(WF)	0.70	100	0.70	100	3.89	100	3.97	100
Meaningfulness	53.3	13	49.3	19	55.5	33	56.8	41
Imagery	65.8	26	67.2	25	57.7	59	62.8	55
Concreteness	61.3	23	65.0	24	55.5	49	60.8	49

Note—HN, high neighborhood density; LN, low neighborhood density; ln(WF), the natural log of Kučera and Francis (1967) word frequency. Meaningfulness, imagery, and concreteness correspond to MEANC, IMAG, and CONC, respectively, as defined by Coltheart (1981), and all have ranges of 100–700. These measures were converted to percentages in the table. % of words, percentage of words for which the measure was available.

was an equal number of words from high-density neighborhoods (three or more neighbors) and from low-density neighborhoods (fewer than three neighbors). Each of these subsets consisted of 140 words. Of these, 84 were five-letter, 42 were six-letter, and 14 were seven-letter words (see the Appendix).² Some words were from the same neighborhood, which could affect results due to orthographic study list category effects or new–old similarity effects (see, e.g., Heathcote, 2003), but such occurrences were relatively rare and had a negligible influence.³

Glanzer and Adams (1985) found that mirror effects usually occurred for manipulations of a number of semantic variables, including concreteness, imageability, and meaningfulness (see their Table 6). Table 1 lists the means for each group of words on these measures, taken from the MRC database. The words were chosen primarily to control frequency and density, so ratings for other properties were available for only some of the words in each set. The high-frequency words are slightly more meaningful, but they have lower imagery and concreteness than the low-frequency words. Differences as a function of density are even smaller, with high-density words having slightly greater imagery and concreteness than low-density words. Given that these differences between word sets are small—much smaller than the differences usually associated with demonstrations of a mirror effect for these variables—it is unlikely that they would confound our frequency and density manipulations.

The participants in both experiments rated their decision confidence and also indicated if a test word was new or old. In their Experiments 3 and 4, Glanzer and Adams (1990) collected confidence ratings in a 2×2 design similar to ours, except that frequency was crossed with concreteness. Typically, accuracy for concrete words is higher than accuracy for abstract words due to a mirror effect (Glanzer & Adams, 1985). Overall, Glanzer and Adams (1990) found separate mirror effects of both frequency and concreteness, but the concreteness effect was weaker for high-frequency than for low-frequency words. In particular, in their Experiment 3 they failed to find the mirror effect for concreteness in HRs for high-frequency words. In their Experiment 4, in which a concreteness encoding task was used, the strength of the overall concreteness effect was increased and the mirror effect was obtained in HRs for high-frequency words, although it was only weak. In both experiments, larger mirror effects for both frequency and concreteness were found in average response confidence, suggesting that it provides a more sensitive measure. Hence, we examined response confidence as well as $p(\text{Old})$ results. Response confidence was also used to construct ROC curves, but reporting of these analyses is postponed until the General Discussion.

EXPERIMENT 1

Method

Participants. Sixteen first-year psychology students at the University of Newcastle participated in the experiment for course credit. Participation was voluntary.

Procedure. For each participant, the 560 words were randomly allocated to one of seven study test lists of equal length, with each list containing an equal number of items from each word set. For each list, half of the words from each of the four word sets were randomly allocated to the study task, with the remaining half used as new words during test. Separate sets of words were utilized as practice and buffer items. Buffer items were presented in the first and the last four study positions. The practice items were 80 words, 20 from each word category. Buffer items were drawn from a set of 64 five- to seven-letter words that had no neighbors.

Testing was undertaken on a computer that presented the stimuli and recorded the participants' responses using a program written in Turbo Pascal 6.0 with millisecond accurate timing (Heathcote, 1988). Stimuli were displayed in lowercase at the center of a 17-in. monitor and presented in white text on a black background. Responses were made using two three-button mouse input devices connected to an IO card in the computer.

The experimental session lasted 1 h. The participants were tested individually in a cubicle that provided a distraction-free environment. They were informed that they would view lists of words and would subsequently be presented with a list containing both old and new words. The participants were instructed to indicate whether each word presented at test was old or new and to simultaneously rate their confidence on a three-point scale.

The experimental session consisted of eight study/test cycles, the first of which was practice. Each study task involved the presentation of 40 words, 10 from each word category, in a random order. In addition, 8 buffer words were presented in each list, randomly allocated to either one of the first four or one of the last four study positions. Each word was presented on the screen for 1,000 msec with a 300-msec interstimulus interval. The participants then completed a distractor task consisting of 10 single-digit multiplication and addition problems. The participants were given two possible answers and required to indicate which answer was correct. At the completion of the distractor task, the participants began the test phase by pressing all six mouse buttons at once.

In the test phase, the studied old words and an equal number of unstudied new words were presented one at a time in a random order. One of the eight studied buffer words was randomly selected to appear in the recognition memory test, at a random position, and the responses for this item were not recorded. The participants indicated whether each item was old or new using a three-point confidence rating scale for each option. The six possible responses were displayed at the bottom of the test screen with the labels *sure old*, *probably old*, *possibly old*, *possibly new*, *probably new*, and *sure new*. Response keys for each confidence rating were the three keys on the left mouse for "new" responses and the three keys on the right mouse for "old" responses, to be pressed using the left and right hands, respectively. If the participant did not respond within 5 sec, the test word was removed from the screen. The participant initiated the next test by pressing the six buttons at once.

The only form of feedback during the experiment occurred immediately after the completion of the practice task. The participants were presented with a screen displaying the frequency with which they used each of the six response categories during the recognition memory test. They were reminded to make use of all six response categories in the subsequent recognition memory tests.

Results

All inferential test results reported are described as significant when $p < .05$. Over all participants, six responses were made in under 200 msec and an additional 29 responses were made in over 5 sec, and so were not recorded. These responses made up 0.39% of the overall number of responses and were removed from further analyses. In order to check for speed–accuracy trade-off

at test, an ANOVA was performed on mean response time (RT) using as factors the word variable (new vs. old) and the response variables (confidence and accuracy). The design was slightly unbalanced because 3.6% of the responses were missing, most of them high-confidence “new” responses to old test items and high-confidence “old” responses to new test items. The missing values were replaced using an additive subjects effect model (i.e., a missing cell mean of an individual participant was replaced by the mean over participants for that cell in the unbalanced design plus the difference between the grand mean RT and the individual participant’s mean RT).

Only response variables showed significant effects, indicating that speed–accuracy trade-off does not confound the following analyses. Confidence had the strongest main effect [$F(2,30) = 18.6$, $MS_e = 360,257$, $p < .001$], with mean RT decreasing from 1.6 to 1.3 sec as confidence increased. “New” responses were about 60 msec slower than “old” responses [$F(1,15) = 6.4$, $MS_e = 96,425$, $p = .023$], and incorrect responses were about 60 msec slower than correct responses [$F(1,15) = 6.18$, $MS_e = 38,932$, $p = .001$].

Table 2 reports accuracy⁴ as measured by d' . Both frequency and density produced significant main effects [$F(1,15) = 13.4$, $MS_e = 0.110$, $p = .002$, and $F(1,15) = 9.14$, $MS_e = 0.037$, $p = .009$, respectively], but they did not interact [$F(1,15) = 1.9$, $MS_e = 0.058$, $p = .19$]. The main effect of frequency on accuracy was twice as large as the density main effect (d' s of approximately 0.3 and 0.15, respectively), with responses to both low-frequency and low-density words being more accurate than responses to their high-frequency and high-density counterparts. Hence, it appears that our manipulation of density was sufficient to affect recognition memory and that, in consistency with Estes and Maddox’s (2002) conjecture, accuracy was lower for words with higher word likeness as measured by density.

Table 3 shows the results for FAR and HR [i.e., $p(\text{Old})$] and mean confidence (R). Confidence was scored on a scale of 1 (*sure new*) to 6 (*sure old*) and averaged separately for new and old items of each word type. For HR, the main effect of frequency was significant [$F(1,15) = 7.58$, $MS_e = 0.0055$, $p = .015$], with low-frequency words (0.74) producing more hits than high-frequency words (0.69). For FAR, the main effect of frequency was also significant [$F(1,15) = 5.49$, $MS_e = 0.0048$, $p = .033$],

with low-frequency words producing a smaller proportion of false alarms (.23) than high-frequency words (.27). Similar results were found for confidence ratings. The main effect for old words was significant [$F(1,15) = 15.6$, $MS_e = 0.085$, $p = .001$], with means of 4.65 for low-frequency words and 4.35 for high-frequency words. The main effect for new words was also significant [$F(1,15) = 16.8$, $MS_e = 0.042$, $p = .001$], with means of 2.5 for low-frequency and 2.7 for high-frequency words. Hence, frequency produced a mirror effect in both “old” response probability and mean confidence.

A mirror effect of density was also found in mean confidence ratings. The main effect of density was significant for old words [$F(1,15) = 5.23$, $MS_e = 0.035$, $p = .037$], with means of 4.55 for low-density and 4.45 for high-density words, and for new words [$F(1,15) = 11.5$, $MS_e = 0.021$, $p = .004$], with means of 2.5 for low-density and 2.65 for high-density words. Word frequency and density interacted for new words [$F(1,15) = 5.68$, $MS_e = 0.013$, $p = .031$] due to a weaker density effect for high- than for low-frequency words, whereas there was no interaction for old words ($F < 1$). Although density produced a mirror pattern in $p(\text{Old})$, only the main effect on HR was significant [$F(1,15) = 9.14$, $MS_e = 0.0017$, $p = .006$], with $p(\text{Old})$ for low-density words (.74) greater than $p(\text{Old})$ for high-density words (.70). For FAR, $p(\text{Old})$ for low-density words (.24) was less than $p(\text{Old})$ for high-density words (.25), but neither the main effect of density ($F < 1$) nor its interaction with frequency [$F(1,15) = 1.78$, $MS_e = 0.002$, $p = .202$] achieved significance.

Discussion

We found the same pattern of frequency effects when density was controlled that Malmberg et al. (2002) found when letter frequency was controlled: significantly lower HRs and higher FARs, and significantly greater accuracy for low-frequency than for high-frequency words. Hence, the frequency mirror effect does not appear to be due solely to word likeness as measured by the frequency of either the lowest or the highest order substrings of words. Like Malmberg et al., we obtained a mirror effect pattern for our measure of word likeness. However, although in both studies the effect on accuracy was significant, their letter frequency effect was reliable for FARs but not HRs, whereas our density effect was reliable for HRs but not FARs. The density effect for new items was, however, reliable in average response confidence, supporting the assertion that this measure is more sensitive than $p(\text{Old})$.

Failure of the mirror effect due to density for new items was restricted to high-frequency words, for which FARs were virtually equivalent (.27). For low-frequency words, in contrast, the FAR for low-density words (.215) was less than the FAR for high-frequency words (.24). Response confidence results were consistent in that the overall significant main effect of density for new words was tempered by a significant interaction with frequency due to a weaker effect for high- than for low-frequency words. A weaker effect of density for high- than for low-frequency

Table 2
Accuracy and z -ROC Measures for Each Word Type

Experiment		High Frequency		Low Frequency	
		HN	LN	HN	LN
1	d'	1.151	1.213	1.372	1.600
	Intercept	0.917	1.093	1.124	1.336
	Slope	0.672	0.740	0.652	0.644
2	d'	1.515	1.776	2.203	2.285
	Intercept	1.315	1.533	1.714	2.013
	Slope	0.701	0.734	0.677	0.718

Note—HN, high neighborhood density; LN, low neighborhood density.

Table 3
Mean Confidence (*R*) on a Scale of 1–6 for “Sure New” to “Sure Old” and
Probability of an “Old” Response [*p*(Old)] for New (i.e., FAR)
and Old (i.e., HR) Items for Each Word Type

Data Set	Measure	New Items				Old Items			
		LF		HF		HF		LF	
		LN	HN	LN	HN	HN	LN	HN	LN
Experiment 1	<i>p</i> (Old)	.215	.240	.270	.266	.679	.708	.726	.763
	<i>R</i>	2.39	2.58	2.67	2.72	4.32	4.38	4.56	4.71
Experiment 2	<i>p</i> (Old)	.146	.176	.196	.240	.773	.804	.844	.870
	<i>R</i>	2.14	2.34	2.43	2.62	4.62	4.76	5.02	5.15

Note—HF, high frequency; LF, low frequency; HN, high neighborhood density; LN, low neighborhood density.

words is consistent with results from the lexical memory literature. Andrews (1989) found a reliable facilitatory effect of density on lexical decision task RTs (i.e., RTs for high-density words were shorter than those for low-density words), but only for low-frequency words, a result that was replicated in 21 of the 25 studies reviewed by Andrews (1997). Thus, increased density appears to facilitate lexical memory, but only for low-frequency words. The results of Experiment 1 suggest a similar conclusion for episodic recognition, except that the effect of density is inhibitory. However, density did have a significant effect on HRs, and, although it was slightly larger for low- than for high-frequency words, the corresponding interaction was not reliable.

EXPERIMENT 2

The aim of Experiment 2 was to resolve the ambiguity in the results for density in Experiment 1 by increasing the overall effect of density. This was achieved by using lexical decision as the study task. Lexical decision is an effective means of encoding for episodic recognition memory that has been found to increase the magnitude of the frequency effect in comparison with that achieved by other encoding tasks (Hilford, Glanzer, & Kim, 1997; Hoshino, 1991; Joordens & Becker, 1997). It seems likely that the emphasis on encoding orthographic features in the lexical decision task would also increase the effect of density on episodic recognition, and so would help determine whether or not our failure to find a reliable effect of density on FARs for high-frequency words in Experiment 1 was a result of a weak overall effect size. Accuracy and RT results from the lexical decision task also allow a direct comparison between the effects of our density manipulation on lexical and episodic memory in the same group of participants.

Method

All aspects of Experiment 2 were the same as for those of Experiment 1, except that lexical decision was used for study and all responses in the recognition memory task were recorded, including those taking longer than 5 sec. Each lexical decision study list required responses to 80 words and an equal number of nonwords, with stimuli drawn at random without replacement from the four word and two nonword sets. The display of each lexical decision stimulus was terminated by a response. Non-

words were selected from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002) and consisted of orthographically legal bigrams and bodies. Nonwords were matched with words for density and number of letters; half of the nonwords were high density (three or more neighbors, mean density = 4.75), and half were low density (fewer than three neighbors, mean density = 1.43). Twenty-three students at the University of Newcastle participated in the experiment.

Results

No study or test responses were made in under 200 msec, and there was no time limit on responses, so no responses were omitted from the following analyses. Table 4 shows the RT and mean percent correct for lexical decision. Responses to high-density nonwords were both 2.5% more error prone [$t(22) = 3.67$, $SE = 0.007$, $p = .001$] and 95 msec slower [$t(22) = 7.18$, $SE = 13.2$, $p < .001$] than responses to low-density nonwords. However, for words, density did not have a significant main effect and did not interact with frequency in either percent correct or RT. In contrast, responses to high-frequency words were both 8.9% less error prone and 147 msec faster than responses to low-frequency words [$F(1,22) = 29.4$, $MS_e = 0.006$, $p < .001$ and $F(1,22) = 77.9$, $MS_e = 6,365$, $p < .001$, respectively].

As in Experiment 1, there was no evidence that a speed–accuracy trade-off during recognition memory tests confounded comparisons between word types,⁵ with response factors having effects on episodic test RT similar to those found in Experiment 1. Both frequency and density produced significant main effects on d' [$F(1,22) = 80.7$, $MS_e = 0.074$, $p < .001$ and $F(1,22) = 20.1$, $MS_e = 0.075$, $p < .001$, respectively] but did not interact ($F < 1$). Both main effects increased relative to those of Experi-

Table 4
Mean Response Times (in Milliseconds) and Percent Correct in
the Lexical Decision Task

Data Set	Mean RT		Percent Correct	
	LN	HN	LN	HN
Nonwords	1,069	1,164	97.3	94.8
Low-frequency words	1,035	1,051	90.0	90.2
High-frequency words	869	866	99.0	98.9

Note—HN, high neighborhood density; LN, low neighborhood density.

ment 1, but the density main effect remained at about half the size of the frequency main effect ($d' = 0.5$ and 0.26 , respectively; see Table 2).

Word frequency had a significant main effect on FAR [$F(1,22) = 26.6$, $MS_e = 0.003$, $p < .001$], with a smaller mean for low-frequency (.16) than for high-frequency (.22) words, and a significant main effect on HR [$F(1,22) = 20.7$, $MS_e = 0.005$, $p < .001$], with a larger mean for low-frequency (.86) than for high-frequency (.79) words. Density had a significant main effect on FAR [$F(1,22) = 20.9$, $MS_e = 0.003$, $p < .001$], with a smaller mean for low-frequency (.17) than for high-frequency (.21) words, and a significant main effect on HR [$F(1,22) = 7.36$, $MS_e = 0.003$, $p = .013$], with a larger mean for low-density (.84) than for high-density (.81) words. No interactions between frequency and density were significant ($F < 1$ for both HR and FAR).

Reliable mirror effects were also found in mean confidence ratings for both frequency and density. Word frequency and density produced significant main effects for new words [$F(1,22) = 60.6$, $MS_e = 0.031$, $p < .001$, and $F(1,22) = 27.6$, $MS_e = 0.032$, $p < .001$, respectively] and for old words [$F(1,22) = 45.4$, $MS_e = 0.081$, $p < .001$, and $F(1,22) = 16.7$, $MS_e = 0.025$, $p < .001$, respectively], with no interactions ($F < 1$ for both FAR and HR). For new words, the mean response confidence rating for low-frequency words (2.25) was smaller than that for high-frequency words (2.5), and the mean confidence rating for low-density words (2.3) was smaller than that for high-density words (2.5). For old words, the mean confidence rating for low-frequency words (5.1) was greater than the mean for high-frequency words (4.7), and the mean for low-density words (5.0) was greater than that for high-density words (4.8).

Discussion

Lexical decision was an effective study task that increased overall accuracy relative to the free study task in Experiment 1. The sizes of both the frequency and density effects were proportionately increased, with the effect of density on accuracy becoming almost equivalent to the frequency effect in Experiment 1. In consistency with the hypothesis that our failure to find reliable results for some aspects of the density mirror effect was due to a small effect size in Experiment 1, the increased density effect in Experiment 2 was associated with reliable differences in the mirror order for HR and FAR and for response confidence. Density and frequency effects in both measures were close to additive, as is indicated by small and unreliable interaction effects. Hence, the results of Experiment 2 confirm the conclusion from Experiment 1 that frequency produces a mirror effect in episodic recognition when word likeness is controlled. Experiment 2 extends the results from Experiment 1 by showing that word likeness, as measured by density, produces a mirror effect in episodic recognition for both high-frequency and low-frequency words. This contrasts with the usual finding in the lexical memory literature that density affects lexi-

cal decision and naming performance for low-frequency words only (Andrews, 1997).

The lexical decision results from study in Experiment 2 allow us to compare the relative effects of our word sets on episodic and lexical memory. Frequency had a large facilitatory effect on lexical memory, with responses to high-frequency words in the lexical decision task almost 10% more accurate and 150 msec faster than responses to low-frequency words. The longer RT for low-frequency words is consistent with greater attention to low- than to high-frequency words during study. Malmberg and Nelson (2003) proposed that the episodic advantage for low-frequency words occurs because these words attract more attentional resources during perceptual identification and lexical access processes that occur early in study. This proposal is consistent with our finding that the frequency effect increased when the lexical decision task was used for study, since this task emphasizes perceptual identification and lexical access processes.

Although our manipulation of density was smaller than is usual in the lexical memory literature, it did produce a reliable inhibitory effect for nonwords in lexical decision, with responses to low-density words 2.5% more accurate and almost 100 msec faster than responses to high-density words. Despite being of the same magnitude as the nonword density manipulation, the density manipulation for words had no effect on lexical decision: Accuracy and RT were virtually identical for low-density and high-density words, even when those words were of low frequency. Despite this null effect on lexical memory, density had a clear inhibitory effect on episodic memory.

The increased episodic density effects in Experiment 2 in comparison with those of Experiment 1 are consistent with the assumption that low-density words, like low-frequency words, attract more attention during perceptual encoding and lexical access processes, resulting in better encoding than high-density words. However, if attention is indexed by lexical decision RT, the lexical decision results are inconsistent with the assumption that low-density words attract more attention. A possible explanation is that RT for lexical decision depends on more than identification. Some lexical models (see, e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Grainger & Jacobs, 1996) assume that lexical decisions can be made on the basis of criteria placed on two processes: identification through competitive selection of an individual lexical representation and the total activation of the lexicon. Identification is slowed by increased density as a result of stronger competition, but total activation increases with density as a result of partial activation of more lexical representations, speeding responses to high-density words. It is possible that our participants, knowing that their memory would be tested later, emphasized the identification process and thus eliminated the small advantage for high- in comparison with low-density words usually found with low-frequency words in lexical decision.

Our results, like those of Malmberg et al. (2002), show that increased word-part frequency, as well as increased

whole-word frequency, is detrimental to episodic recognition accuracy. The fact that frequency differences in both whole words and word parts result in an episodic mirror effect suggests that they may act through a common mechanism. For example, in Shiffrin and Steyvers's (1997) REM (retrieving-effectively-from-memory) theory, more frequent features are less distinctive. In Glanzer and Adams's (1990) attention likelihood theory (ALT), not only words that are more frequent overall, but also words with more frequent parts, may attract less attention at study. In both cases, a likelihood transformation causes these differences to affect both "new" and "old" responses, producing a mirror effect. In dual-process theories (see, e.g., Joordens & Hockley, 2000), greater frequency may produce increased familiarity in new words but reduce retrieval for old words, perhaps because higher frequency features are linked to more contexts. Although our results do not discriminate between these mechanisms, the fact that both frequency effects and density effects were increased proportionally by an encoding task that emphasized orthographic processing is consistent with a common mechanism that is sensitive to frequency of experience with orthographic features.

GENERAL DISCUSSION

Our results do not support Estes and Maddox's (2002) hypothesis that the word frequency mirror effect is due to correlated differences in word likeness. Our finding that frequency produces a mirror effect when density is controlled is added to a list of recent demonstrations that the word frequency effect is not due to correlated characteristics such as letter frequency (Malmberg et al., 2002), the normative frequency of contexts in which a word occurs (Steyvers & Malmberg, 2003), age of acquisition (Dewhurst, Hitch, & Barry, 1998), or the richness of associative connectivity (Nelson, Zhang, & McKinney, 2001).

Estes and Maddox (2002) formulated their hypothesis to accommodate findings of a parallel effect of short-term familiarization (see also Chalmers & Humphreys, 1998; Maddox & Estes, 1997), on the basis of the assumption that familiarization training is an experimental analogue of normative word frequency. The latter assumption may, however, be wrong, at least for short-term training. When Reder, Angstadt, Cary, Erickson, and Ayers (2002) gave nonwords much longer familiarization training, which is more representative of natural language experience (up to 360 study and recall tests distributed over 5 weeks), they initially found a parallel effect of different levels of training on episodic recognition, but later in training they found a mirror effect. These results seem to indicate that natural language experience has two distinct effects on episodic recognition: a short-term parallel effect and a longer term mirror effect.

Our results do support Estes and Maddox's (2002) hypothesis that orthographic word likeness produces a mirror effect, with higher levels of word likeness (as measured by neighborhood density) associated with reduced episodic

recognition. The strength of the density effect was dependent on study encoding, with a study task that increased emphasis on encoding orthographic features (lexical decision) increasing both the density and frequency effects relative to a free-study task. In both cases, relatively brief study times (on the order of 1 sec) may also have emphasized both density and frequency effects. Malmberg and Nelson (2003) found that extending study time to several seconds did not increase the frequency effect, which they attributed to equal attention given to low-frequency and high-frequency words during the semantic processing that occurred after identification. Since low-density and high-density words are also unlikely to differ in semantic processing, longer study durations are unlikely to increase the density effect and may even decrease it if participants emphasize semantic features at test.

We used neighborhood density as a measure of orthographic word likeness because it has been widely studied in the lexical memory literature. Although it is a measure of the frequency of a word's highest order parts, it has limitations as a comprehensive measure of word likeness and part frequency. For example, density is length specific, so it does not measure the frequency of strings across all word lengths, does not weight substring frequency by the frequency of words in which the substrings occur, and does not take into account the frequency of lower order substrings, including that of salient word parts, such as morphemes and syllables. In the following section, we examine effects due to the smallest parts of words—letters and bigrams—in our data.

Letter and Bigram Frequency

Some caution should be exercised in concluding that recognition memory for a word is directly affected by the number of neighbors that it has. Word manipulations are often open to confounding due to the highly correlated nature of many word characteristics. As is shown in Table 1, confounding of our word sets due to semantic attributes such as concreteness is unlikely. In lexical memory, Grainger (1990) suggested that the facilitatory density effects found by Andrews (1989) might have been mediated by bigram frequency. Bigram frequency refers to the frequency of occurrence of adjacent pairs of letters in the language and is positively correlated with density. Andrews (1992) subsequently replicated her 1989 study controlling for bigram frequency. The facilitatory effect of density for low-frequency words was still observed, in both lexical decision and naming tasks. Furthermore, when density was held constant, there was no effect of bigram frequency in lexical decision. In episodic memory, Zechmeister's (1972) mirror effect due to differences in a word's orthographic distinctiveness, as rated by participants in Zechmeister (1969), may have been due to bigram frequency differences. Zechmeister (1969) reported that words rated as distinctive, such as *gnome* and *slyph*, had lower average bigram frequencies. Because our low-density words were chosen to have at least one neighbor, they tended not to contain the rare bigrams evident in

Zechmeister's (1972) distinctive words, but the possibility of confounding by bigram frequency remains.

In order to determine whether bigram frequency might be mediating density effects, we calculated bigram frequency for our word sets based on the corpus of five- to seven-letter words in Coltheart's (1981) MRC Dictionary (excluding words containing punctuation) with Kučera and Francis's (1967) frequencies greater than zero. This corpus consists of 2,028 five-letter words, 2,645 six-letter words, and 2,687 seven-letter words. We computed the number of times each bigram occurred at each position within each word length set and converted these counts to percentages of the total count to form a measure of type bigram frequency. We also computed a measure of token bigram frequency by multiplying the occurrences of each bigram in each word by the word's frequency count. The resulting token counts were converted to percentages of the total number of token occurrences. For each word in the experimental word sets, we calculated the mean type and token percentages over each position. Table 5 shows the resulting means for each word set. Both type and token bigram frequencies are confounded with density. Word frequency is fairly independent of type bigram frequency but more strongly confounded with token bigram frequency.

In order to remove the confounding of density by type and token bigram frequency, we selected subsets of responses for each participant in Experiments 1 and 2 (see also Steyvers & Malmberg, 2003). The subsets were created by selecting equal numbers of responses to new and old test words with the highest bigram frequencies for low-density words and the lowest bigram frequencies for high-density words.⁶ Table 5 shows the statistics of the resulting type- and token-equated subsets. The density and frequency manipulations were only slightly weakened in these subsets, so if their effects on recognition memory

are not mediated by bigram frequency, the same pattern of results found in the full set should be observed in each equated set.

Tables 6 and 7 report recognition memory results for the combined data from both experiments for all responses and for the type- and token-equated subsets. For d' , the main effects of both frequency and density were significant for all responses and for the type- and token-equated subsets, but no interactions between density and frequency approached significance. Similarly, for FAR and HR, and for mean confidence ratings for new and for old words, no interactions between frequency and density were significant. However, the main effects of frequency and density were significant in all cases, with the exceptions of marginally significant main effects of density for type-equated HR [$F(1,38) = 2.92$, $MS_e = 0.0031$, $p = .096$] and token-equated mean confidence ratings for old words [$F(1,38) = 3.84$, $MS_e = 0.040$, $p = .057$]. Hence, it appears that both frequency and density produce a mirror effect independent of type or token bigram frequency.

In order to check whether letter frequency was confounded with word frequency and density, we calculated mean letter frequency scores for our words using Malmberg et al.'s (2002) methods. Table 5 presents the results for the full word set and the type and token bigram-frequency-equated subsets. In the full set, letter frequency increased very slightly with both density and frequency, by around 0.002 for density and 0.004 for frequency. In comparison, Malmberg et al.'s manipulation of letter frequency was an order of magnitude larger (0.04 on average). The differences in letter frequency were further attenuated for the type- and token-equated subsets, so it can be concluded that the frequency and density effects were not mediated by letter frequency.

Although differences in letter and bigram frequency do not appear to have caused our density effects, they may have been mediated by the frequency of higher order word parts such as those constituting the "Wickelfeature" representation used in Seidenberg and McClelland's (1989) lexical model or the onset-vowel-coda representation used in Plaut, McClelland, Seidenberg, and Patterson's (1996) lexical model. Regardless of the exact source of the density effects that we observed, our results indicate that recognition memory is sensitive to higher order word-part frequency effects at some level, particularly when study emphasizes word identification processes. Hence, both lexical and episodic memory models may be informed by further episodic recognition studies in which the frequency of the higher order word parts hypothesized to make up word representations in lexical memory are controlled and manipulated.

Multiple Mirror Effects

Up until now, we have considered mainly the mirror effects due to density and frequency separately. Glanzer and Adams (1990) showed that ALT predicts the full joint order of conditions in a 2×2 design when the effect of one mirror variable is stronger than that of another. In our case, the effect of frequency on accuracy was approxi-

Table 5
Type and Token Mean Position-Specific Bigram Frequencies and Letter Frequencies for All Words Used in Experiments 1 and 2, and for Type- and Token-Equated Subsets of These Words, Along With the Size of Each Subset for Each Participant and Their Mean Densities and Natural Log Word Frequencies [$\ln(WF)$]

Word Set	Measure	Low Frequency		High Frequency	
		LN	HN	LN	HN
All words	Type	1.05	1.71	1.19	1.78
	Token	0.99	1.65	1.42	2.15
	Letter freq.	0.072	0.074	0.076	0.078
	Size	100	102	94	92
Type equated	Density	1.34	4.71	1.53	4.61
	$\ln(WF)$	0.69	0.73	3.95	4.03
	Type	1.25	1.25	1.26	1.26
	Token	1.21	1.15	1.49	1.53
	Letter freq.	0.075	0.074	0.078	0.079
Token equated	Number	94	122	94	84
	Density	1.33	4.73	1.46	4.72
	$\ln(WF)$	0.71	0.70	3.94	3.84
	Type	1.25	1.43	1.10	1.30
	Token	1.30	1.30	1.29	1.29
	Letter freq.	0.076	0.075	0.078	0.077

Note—HN, high neighborhood density; LN, low neighborhood density.

Table 6
Mean d' for Experiments 1 and 2 Data Combined and for the Combined Data Equated on Type and Token Bigram Frequency

Data Set	High Frequency		Low Frequency	
	HN	LN	HN	LN
All	1.366	1.545	1.756	2.004
Type	1.391	1.499	1.765	1.963
Token	1.414	1.583	1.724	1.930

Note—HN, high neighborhood density; LN, low neighborhood density.

mately twice that of density. To provide a simplified illustration (which corresponds approximately to the observed values for all data combined in Table 6), suppose that $d' = 0.4$ and 0.2 for the frequency and density effects, respectively, and that for the least accurate condition (high frequency [HF] and high neighborhood density [LN]) $d'(\text{HFHN}) = 1.4$. Relative to this condition, for the condition with HF and low neighborhood density (LN), accuracy is increased by 0.2 , so that $d'(\text{HFLN}) = 1.6$, whereas with low frequency (LF) accuracy is increased by 0.4 , so that $d'(\text{LFHN}) = 1.8$. If density and frequency effects are additive, decreasing both increases accuracy by $0.2 + 0.4 = 0.6$, so that $d'(\text{LFLN}) = 2$. ALT does not predict exactly additive effects, but it does predict that the order $\text{HFHN} < \text{HFLN} < \text{LFHN} < \text{LFLN}$ holds for accuracy. The columns in Tables 2 and 6 are in this order, and, as predicted, all results increase from left to right.

ALT also predicts the full order of separate measures for new and old items [i.e., $p(\text{Old})$ and mean confidence]. The order for old items is the same as that for accuracy, and the order for new items is its mirror: $\text{LFLN} < \text{LFHN} < \text{HFLN} < \text{HFHN}$. Table 3 shows that the predicted order was obtained for both $p(\text{Old})$ and mean response confidence in Experiment 2, and for mean response confidence in Experiment 1, and only one of the eight predicted sequential orders was violated for $p(\text{Old})$ in Experiment 1. Table 7 shows that for the combined data the full order was obtained in $p(\text{Old})$ for all responses, and for responses to the type and token bigram-equated word sets.

Our multiple mirror effect results are a stronger confirmation of ALT's predictions than was obtained by Glanzer and Adams (1990) for their factorial manipulation of frequency and concreteness. In their Experiment 4, they found only a partial order for the old items in their Experi-

ment 3 and a full order, but one with a very small difference, for high-frequency concrete versus low-frequency abstract FARs. ALT predicts a full mirror order for the joint effects of two mirror variables because it attributes each mirror effect to a single common underlying factor. This prediction is not necessarily unique to ALT. Even theories that attribute mirror effects for new and old items to separate mechanisms (e.g., familiarity and recollection in dual-process theories) could accommodate a multiple mirror effect, but this would seem to require the assumption that both mechanisms are affected by a common underlying factor. Hence, at least at some level, the present results support a single-factor account of mirror effects.

Given that ALT potentially provides a unified explanation of our results, we tested its more detailed predictions about the distribution of response confidence using slope and intercept measures derived from z -ROC analysis. ALT predicts linear z -ROC functions, and so it is a special case of the normal unequal variance signal detection model. We fit the latter model using simultaneous maximum likelihood estimation over all conditions (Kijewski, Swenson, & Judy, 1989; Sheu & Heathcote, 2001), but separately for each participant. The analysis assumes that decision criteria do not vary as a function of word type, an assumption that is likely true given our within-list manipulation of word type (see Stretch & Wixted, 1998). Figure 1 illustrates the fit of the unequal variance normal model by plotting the (predicted) average model and the (observed) average deviations from the average model over participants.⁷ In consistency with ALT's predictions, linear z -ROC functions provided an excellent fit to participant data. Only 1 participant in each experiment had a significant misfit according to a χ^2 test [$\chi^2(21) = 33.3, p = .043$ for Experiment 1 and $\chi^2(21) = 34.97, p = .028$ for Experiment 2], and the total χ^2 over participants did not approach significance in either experiment [$\chi^2(336) = 346.5, p = .67$ for Experiment 1 and $\chi^2(483) = 462.1, p = .25$ for Experiment 2].

Table 2 presents the intercept and slope estimates from z -ROCs that plot new against old for each word type. The intercept is a measure of accuracy, whereas slope measures the variability of confidence for new relative to old items. Slope is usually less than one (see, e.g., Ratcliff, Sheu, & Gronlund, 1992), indicating greater variability for confidence for old words than for new words. The in-

Table 7
Results for New and Old Items for Each Word Type for Experiments 1 and 2 Data Combined, and for the Combined Data Equated on Type and Token Bigram Frequency

Data Set	Measure	New Items				Old Items			
		LF		HF		HF		LF	
		LN	HN	LN	HN	HN	LN	HN	LN
Experiments 1 and 2	$p(\text{Old})$.174	.202	.226	.251	.735	.764	.795	.826
	R	2.24	2.44	2.53	2.66	4.49	4.60	4.83	4.97
Type equated	$p(\text{Old})$.179	.203	.230	.246	.736	.751	.800	.816
	R	2.29	2.45	2.54	2.62	4.48	4.56	4.84	4.92
Token equated	$p(\text{Old})$.178	.209	.218	.251	.749	.770	.794	.811
	R	2.27	2.46	2.54	2.64	4.53	4.60	4.82	4.88

Note—HF, high frequency; LF, low frequency; HN, high neighborhood density; LN, low neighborhood density.

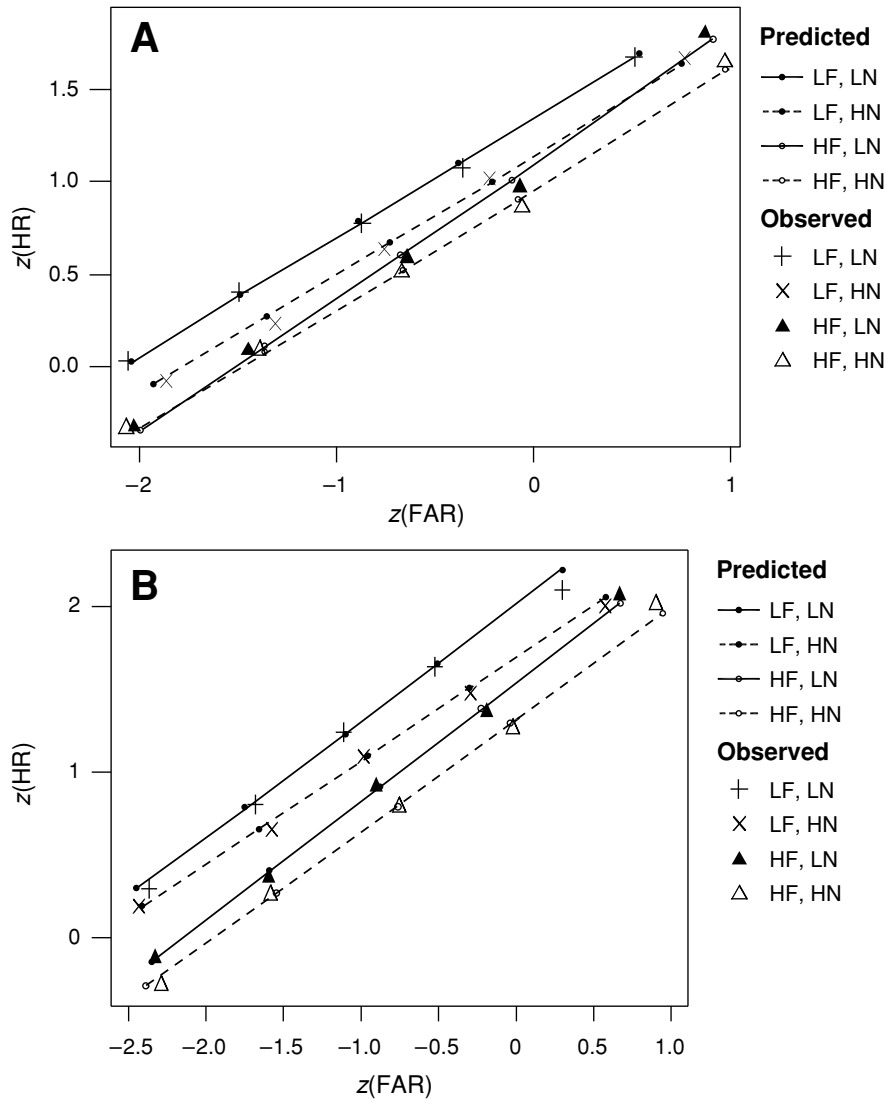


Figure 1. z-ROC plots of the maximum likelihood average unequal variance normal model (predicted) and the average deviations from that model (observed) for (A) Experiment 1 and (B) Experiment 2. HR, hit rate; FAR, false alarm rate; HF, high frequency; LF, low frequency; HN, high neighborhood density; LN, low neighborhood density.

tercept showed the same pattern of results as d' , with significant main effects for both frequency and density for Experiment 1 [$F(1,15) = 7.81, MS_e = 0.104, p = .014$ and $F(1,15) = 32.8, MS_e = 0.018, p < .001$, respectively] and Experiment 2 [$F(1,22) = 41.5, MS_e = 0.107, p < .001$ and $F(1,22) = 17.4, MS_e = 0.089, p < .001$, respectively] but no interaction ($F < 1$ for both experiments). These results are predicted by ALT, as is the full order of intercepts across density and frequency conditions, which ALT predicts to be the same as for d' .

ALT predicts that slope should be greater for conditions with less accuracy, so slope should be greater for high-frequency than low-frequency words and greater for high-density than low-density words. ALT also predicts approximately additive slope effects, so that the full order

of slopes should be $HFHN > HFLN > LFHN > LFLN$. Hence, in contrast to accuracy measures, slope estimates are predicted to decrease from left to right (see Table 2). In consistency with ALT's predictions, the main effect of slope was greater for high- than for low-frequency words in both Experiment 1 (0.71 vs. 0.65, respectively) and Experiment 2 (0.72 vs. 0.70, respectively), but neither effect achieved significance [$F(1,15) = 2.10, MS_e = 0.025, p = .168$ and $F < 1$, respectively]. However, the main effect of slope for density was in the direction opposite that predicted by ALT, with greater slope for low- than for high-density words in both Experiment 1 (0.69 vs. 0.66) and Experiment 2 (0.725 vs. 0.69); again, neither effect was significant [$F < 1$ and $F(1,15) = 1.01, MS_e = 0.03, p = .33$, respectively].

It would be premature to reject ALT on the basis of this one failure to account for density effects on confidence variability, particularly before these effects are replicated using other word sets and before more powerful experiments are able to determine whether or not the slope trends observed in the present experiment are reliable. The predictions that we tested were derived from ALT under the simplifying assumption of common parameters for all members of a class of words. Our low-density words all had 1 or 2 neighbors ($SD = 0.5$), whereas our high-density words had 3 to 10 neighbors ($SD = 1.5$). The difference in variability may explain our failure to find the predicted slope order. However, we note that the difference in log word frequency variability was much more extreme for our low-frequency set ($SD = 0.7$) than for our high-frequency set ($SD = 3.8$), so unless item variability effects are stronger for density than for frequency, a more fundamental modification of ALT may be required.

REFERENCES

- ANDREWS, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 802-814.
- ANDREWS, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 234-254.
- ANDREWS, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, **4**, 439-461.
- BAAYEN, R. H., PIEPENBROCK, R., & VAN RIJN, H. (1993). *The CELEX lexical database* [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.
- BALOTA, D. A., & CHUMBLEY, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 340-357.
- CHALMERS, K. A., & HUMPHREYS, M. S. (1998). Role of generalized and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 610-632.
- COLTHEART, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.
- 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). Hillsdale, NJ: Erlbaum.
- COLTHEART, M., RASTLE, K., PERRY, C., LANGDON, R., & ZIEGLER, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, **108**, 204-256.
- DAVIS, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, **37**, 65-70.
- DENNIS, S., & HUMPHREYS, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, **108**, 452-478.
- DEWHURST, S. A., HITCH, G. J., & BARRY, C. (1998). Separate effects of word frequency and age of acquisition in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 284-298.
- 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, & Cognition*, **28**, 1003-1018.
- 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, & Cognition*, **16**, 5-16.
- GLANZER, M., ADAMS, J. K., IVERSON, G. J., & KIM, K. (1993). The regularities of recognition memory. *Psychological Review*, **100**, 546-567.
- GLANZER, M., & BOWLES, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **2**, 21-31.
- GRAINGER, J. (1990). Word frequency and neighbourhood frequency effects in lexical decision and naming. *Journal of Memory & Language*, **29**, 228-244.
- GRAINGER, J., & JACOBS, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, **103**, 518-565.
- HEATHCOTE, A. (1988). Screen control and timing routines for the IBM microcomputer family using a high-level language. *Behavior Research Methods, Instruments, & Computers*, **20**, 289-297.
- HEATHCOTE, A. (2003). Item recognition memory and the ROC. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1210-1230.
- HILFORD, A., GLANZER, M., & KIM, K. (1997). Encoding, repetition, and the mirror effect in recognition memory: Symmetry in motion. *Memory & Cognition*, **25**, 593-605.
- HOSHINO, Y. (1991). A bias in favor of the positive response to high-frequency words in recognition memory. *Memory & Cognition*, **19**, 607-616.
- JOORDENS, S., & BECKER, S. (1997). The long and short of semantic priming effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 1083-1105.
- JOORDENS, S., & HOCKLEY, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1534-1555.
- KIJEWski, M. F., SWENSSON, R. G., & JUDY, P. F. (1989). Analysis of rating data from multiple-alternative tasks. *Journal of Mathematical Psychology*, **33**, 428-451.
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- MADDOX, W. T., & ESTES, W. K. (1997). Direct and indirect stimulus frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 539-559.
- 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.
- MCCLELLAND, J. L., & CHAPPELL, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition. *Psychological Review*, **105**, 724-760.
- MURDOCK, B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin & Review*, **10**, 570-588.
- NELSON, D. L., ZHANG, N., & MCKINNEY, V. M. (2001). The ties that bind what is known to the recognition of what is new. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1147-1159.
- PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph*, **76**(1, Pt. 2).
- PLAUT, D. C., MCCLELLAND, J. L., SEIDENBERG, M. S., & PATTERSON, K. (1996). Understanding normal and impaired reading: Computational principles in quasi-regular domains. *Psychological Review*, **103**, 56-115.
- RASTLE, K., HARRINGTON, J., & COLTHEART, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, **55A**, 1339-1362.
- RATCLIFF, R., SHEU, C.-F., & GRONLUND, S. (1992). Testing global memory models using ROC curves. *Psychological Review*, **99**, 518-535.
- REDER, L. M., ANDERSON, J. R., & BJORK, R. A. (1974). A semantic interpretation of encoding specificity. *Journal of Experimental Psychology*, **102**, 648-656.
- 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, & Cognition*, **28**, 138-152.
- REDER, L. M., NHOUYVANISVONG, A., SCHUNN, C. D., AYERS, M. S.,

- ANGSTADT, P., & HIRAKI, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 294-320.
- SCHILLING, H. E. H., RAYNER, K., & CHUMBLEY, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, *26*, 1270-1281.
- SEIDENBERG, M. S., & MCCLELLAND, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523-569.
- SHEU, C.-F., & HEATHCOTE, A. (2001). A nonlinear regression approach to estimating signal detection models for rating data. *Behavior Research Methods, Instruments, & Computers*, *33*, 108-114.
- SHIFFRIN, R. M., & STEYVERS, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145-166.
- SIKSTRÖM, S. (2001). The variance theory of the mirror effect in recognition memory. *Psychonomic Bulletin & Review*, *8*, 408-438.
- STEVYERS, M., & MALMBERG, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 760-766.
- STRETCH, V., & WIXTED, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1379-1396.
- WIXTED, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 681-690.
- YONELINAS, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*, 1341-1354.
- ZECHMEISTER, E. B. (1969). Orthographic distinctiveness. *Journal of Verbal Learning & Verbal Behavior*, *8*, 754-761.
- ZECHMEISTER, E. B. (1972). Orthographic distinctiveness as a variable in word recognition. *American Journal of Psychology*, *85*, 425-430.

NOTES

1. Thanks to Ken Forster for supplying Psysoft Version 3.22, which calculates density values based on a corpus of around 10,000 words, including most of the words in the Kučera and Francis (1967) corpus.
2. One word in the low-frequency, low-density condition, *spine*, was incorrectly displayed as *spiny* during the experiments. Since *spiny* has no Kučera and Francis (1967) rating, responses to it were removed from all analyses.
3. Test words with a neighbor in the same test list constituted 3.0% of responses in Experiment 1 and 2.1% of responses in Experiment 2. Removing responses to these words had little effect in either experiment, changing $p(\text{Old})$ by at most 0.006 and on average by less than 0.002.
4. The same pattern of results was obtained with both d' and the high-threshold accuracy measure, $HR - FAR$. The d' measure is reported in order to facilitate comparison with the z -ROC intercept.
5. One interaction with a stimulus factor was significant: the three-way interaction between density, new versus old words, and incorrect versus correct responses [$F(1,22) = 18.0$, $MS_e = 69,893$, $p < .001$]. New-versus-old and incorrect-versus-correct factors interacted because, for correct responses, "old" was faster than "new," but for incorrect responses "old" was slower than "new." The three-way interaction came about because the change from correct to incorrect was greater for high- than for low-density words. This single interaction may have been a Type 1 error and did not appear to confound any further analyses.
6. Because words were randomly allocated as new or old for each participant and responses to some words were occasionally missing, the actual words selected varied slightly from participant to participant. The words that were most frequently selected are noted in the Appendix.
7. The deviation from the model, rather than the average over observed values, is plotted because some values were missing for some participants, usually those with high-confidence errors. The missing values can introduce systematic deviations from linearity that are not representative of deviations in any individual participant. This problem is avoided by plotting the average deviation (Heathcote, 2003).

APPENDIX

Word Sets Used in Experiments 1 and 2

Low-Frequency, Low-Neighborhood-Density Words

adore, ^ agony, * aisle, * align, ^ amber, ^* amiss, antic, * arson, ashen, * astral, ^* avert, * badge, ^* baggy, bathe, ^* befell, ^* beget, ^* behold, * bliss, ^* bogus, bonnet, ^* brassy, ^ brittle, ^* cameo, canal, ^* canon, ^ carob, ^* cavern, ^* clench, ^* confine, ^* cruise, ^* crumble, ^* crypt, defer, ^* defiant, ^* delude, ^* depress, ^* devour, ^ dignify, ^* dingo, ^ divert, ^* droop, ^* dryer, ^* eraser, ^* excite, ^* expel, fateful, ^ felon, ^ flail, ^* fluff, ^ flute, ^* foray, ^* foyer, ^* friar, ^* froth, ^* furor, ^ furrow, ^* fussy, gauze, genie, gleam, ^* goddess, ^* grocer, ^* gurgle, ^* hasten, ^* hockey, ^* hoist, ^* horrid, ^* hubby, husky, idiot, inept, inmate, ^* inset, jealous, ^* jewel, ^ jumper, ^* libel, ^ limber, ^* livid, ^* lotus, lucid, madman, ^* marsh, ^* matron, ^* messy, mourn, ^* naive, ^* nymph, outlaw, pagan, ^* pardon, ^* parrot, ^* pelvis, perish, ^* perky, ^ piping, ^* plaid, ^* plaque, plunder, ^* plush, ^* prowl, ^* puppy, quell, ^* regal, ^* repose, ^* revert, ^* rowdy, ^* rupture, ^* sadist, ^* satin, ^* seller, ^* shady, ^* sheaf, ^* shrug, ^* shunt, ^* sieve, ^* sized, ^* snowy, soggy, spongy, spree, ^* squeak, stocky, swarm, ^* tailor, ^* tardy, ^* tempt, tenor, ^* throne, ^* tilth, ^* torment, ^* toxin, * unruly, ^* vault, * vogue, wallop, ^* wistful, ^* wrapper, ^* yodel, ^*

Low-Frequency, High-Neighborhood-Density Words

abode, ^* alloy, ^* bagged, * banker, * banking, baron, ^* barrow, ^* basil, ^* batch, ^* bearer, bidding, billed, * binge, ^* booted, * bowing, * brake, ^* breach, ^* brink, * broom, ^* brunt, ^* budge, ^* bully, ^* caddy, ^* camper, caste, ^* champ, ^* cheery, ^* chive, ^* chore, ^* chump, ^* cleat, ^* click, ^* cling, ^* clove, ^* cluck, ^* clump, ^* covet, ^* cramp, ^* crank, * crate, * crone, ^* curly, ^* ditty, ^* dolly, ^* drawl, ^* dreamy, ^* dresser, * dummy, ^* fairy, ^* fanning, filler, * flatter, * flick, ^* flint, * float, ^* flown, ^* folder, gander, gaunt, ^* gender, giggle, ^* gouge, ^* grail, ^* greed, ^* halter, harden, * heady, ^* hearth, ^* hedge, ^* hinged, * hither, hobble, ^* hopped, ^* hovel, ^* hurdle, ^* infect, ^* infest, ^* kettle, ^* leaky, ^* liner, * logged, ^* louse, ^ lumpy, ^* madding, madly, ^* matting, mellow, ^* menial, ^* miner, * munch, ^* nasal, ^* packet, ^* pasty, ^* pegging, pickle, ^* pinch, ^* platter, * plume, ^* puddle, ^* rabble, ^* rapping, retch, ^* revel, ^* revise, ^* rubble, ^* rumble, ^* sallow, ^* salve, ^* scalp, ^* scare, ^* scorn, ^* scuffle, ^* sever, shack, ^* shine, ^* shove, ^* silky, ^* sinus, ^* slant, ^* slash, ^* sleet, ^* slipper, * snack, ^* spatter, * spill, ^* spiny, stout, ^* strap, ^* strove, ^* strung, ^* stunk, ^* swine, ^* talker, * tonic, ^* traded, * tripe, ^* wager, ^* warring, wordy, ^* wrack, ^*

APPENDIX (Continued)

High-Frequency, Low-Neighborhood-Density Words

abuse, ^* admit, adopt, advice, ^* alive, ^* among, * anger, ^* angle, appear, arise, ^* around, ^* assume, at-tack, aware, ^* basket, ^* become, * began, ^* breath, ^ breeze, brief, ^* broke, ^* buffer, ^* build, ^* burning, * burst, ^ caught, ^* cause, ^* cellar, ^ chair, ^* change, ^* chief, ^* child, ^* civil, clarify, ^ cloud, ^* coating, * crisis, dance, ^* define, ^* delay, depend, ^* depth, * divine, ^* driver, * effect, * eighth, enjoy, erect, exact, fault, * fifty, floor, ^* formal, ^* frame, ^ fresh, ^* front, ^* gloom, ^* going, ^* guess, * handle, ^* healthy, ^* hence, ^* hid-den, ^* honor, ^* hunter, * inner, ^* joint, ^* juice, ^* landed, ^* learn, ^* lemon, lovely, ^* massive, ^* mature, ^* mayor, ^ million, ^* missing, * month, ^* mostly, ^* motor, ^* muddy, mystery, ^* other, ^* outset, ^ piece, ^* pilot, pistol, ^* plastic, ^ player, ^* posse, ^* precise, ^* prison, ^* product, ^* pulled, ^* queen, ^* quick, ^* quite, ^* quote, ^ range, ^ rapid, ratio, * rebut, remove, ^* reply, ^ reveal, ^* ridge, ^ royal, ^ rural, ^ shallow, ^ sixth, * slope, ^ sorry, ^* spend, ^* spray, ^ squat, stand, ^* status, ^ stern, ^* strain, ^* stream, ^* submit, subtle, ^* swift, talking, * teeth, * throat, throw, * total, ^* trust, ^* union, unite, ^* until, * utter, ^* valley, ^* value, viola, waist, ^ witness, ^* women, ^* yield*

High-Frequency, High-Neighborhood-Density Words

ballet, * barbed, * belly, ^* birth, ^* black, ^* blank, ^* boots, ^* bread, ^* bring, brush, ^* calling, candy, ^* cattle, chart, ^* chase, chest, ^ class, ^* closed, ^* content, couch, ^ count, cried, ^* crown, ^* daily, ^* danger, dining, draft, ^* dream, ^* dress, ^* drill, ^* eating, enter, ^ faced, ^* farmer, faster, finding, finger, firing, fixed, ^* flash, ^* formed, * forth, ^* funny, ^* glaze, ^* grade, ^* grand, guilt, ^* heard, ^ horse, ^ humble, ^* insure, ^* intent, ^* lead-ing, leaning, least, ^ lesson, ^* letter, lighter, lightly, ^* living, locked, * lodge, ^* loose, ^ lover, lunch, ^* lying, ^* maker, making, manner, march, ^* mental, ^* metal, ^* middle, ^* model, ^* naval, ^* nearly, ^* needs, ^* no-tion, ^* older, ^ packing, paint, ^* paper, parker, party, ^* picked, * pitch, ^* place, ^ pocket, ^* pointed, * porch, ^* prime, ^* prove, ^* purse, ^* puzzle, ^* racing, reach, ^ recent, ^ reduce, ^* riding, rough, ^ sailing, saying, shaking, shall, ^* sheep, ^* sheer, shirt, ^* simple, ^* single, * sitting, skill, ^* slide, ^* slight, ^* snake, ^* space, ^* spoke, ^* spring, start, ^* staying, stock, ^* strike, ^* stuff, ^* swear, ^* swept, ^* swing, ^* taken, ^* tense, ^* these, ^ ticket, ^* tired, ^* tower, trace, train, * trying, ^ washed, ^* water, while, ^ whole, ^* winning, would^

^Words usually included in the type bigram-equated set.

*Words usually included in the token bigram-frequency-equated set.

(Manuscript received September 1, 2002;
revision accepted for publication May 26, 2005.)