Journal of Experimental Psychology:
Learning, Memory, and Cognition
2002, Vol. 28, No. 5, 830–842

0278-7393/02/$5.00 DOI: 10.1037//0278-7393.28.5.830

Word Frequency and Receiver Operating Characteristic Curves in Recognition Memory: Evidence for a Dual-Process Interpretation

Jason Arndt and Lynne M. Reder
Carnegie Mellon University

Dual-process models of the word-frequency mirror effect posit that low-frequency words are recollected more often than high-frequency words, producing the hit rate differences in the word-frequency effect, whereas high-frequency words are more familiar, producing the false-alarm-rate differences. In this pair of experiments, the authors demonstrate that the analysis of receiver operating characteristic (ROC) curves provides critical information in support of this interpretation. Specifically, when participants were required to discriminate between studied nouns and their plurality reversed complements, the ROC curve was accurately described by a threshold model that is consistent with recollection-based recognition. Further, the plurality discrimination ROC curves showed characteristics consistent with the interpretation that participants recollected low-frequency items more than high-frequency items.

One of the most replicable empirical results in the recognition memory literature is the word-frequency effect. The word-frequency effect is the finding that low-frequency words show superior recognition relative to high-frequency words, both in terms of a higher hit rate and a lower false-alarm rate (Glanzer & Bowles, 1976; Gorman, 1961). Such a pattern of results (i.e., that higher hit-rates are often accompanied by lower false-alarm rates) has been dubbed the *mirror effect* by Glanzer and colleagues (Glanzer & Adams, 1985; Glanzer, Adams, Iverson, & Kim, 1993). Although the mirror effect is more general than manipulations of word frequency (e.g., encoding manipulations such as increased study time also produce mirror effects; Hirshman, 1995; Ratcliff, Clark, & Shiffrin, 1990; Stretch & Wixted, 1998), accounting for the word-frequency mirror effect has proven especially difficult for many theories of recognition memory. In particular, the word-frequency mirror effect has been difficult for global matching models (Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1982) to explain. The reason for this is simple: Models that assume that a single strength dimension underlies recognition memory must explain why low-frequency words show lower levels of memory strength than high-frequency words when they are unstudied but higher levels of memory strength when they are studied.

In light of the difficulty that strength-based single-process models have faced in explaining the mirror effect in general, and the word frequency effect in particular, new theories have been proposed that account for mirror effects, including the word-frequency effect. Some of these theories maintain the assumption that a single process underlies recognition memory performance (Benjamin, Bjork, & Hirshman, 1998; Glanzer et al., 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), whereas others rely on the assumption that two processes contribute to recognition memory performance (Joordens & Hockley, 2000; Reder et al., 2000). Accordingly, both of these approaches have proven successful in accounting for the word-frequency effect, with the specific details of the explanation differing between models.

Single-process explanations of the mirror effect are based on factors (e.g., memory strength, word frequency) that affect the separation of the underlying strength distributions. The greater separation of the distributions underlying recognition memory, when coupled with a decision rule that maximizes memory performance, produces the pattern of greater hits and lower false alarms for the more memorable item class. In model terms, the decision rule is instantiated as the computation of a likelihood ratio comparing the evidence in memory that an item is old with the evidence that the item is new (Benjamin et al., 1998; Glanzer et al., 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). If there is more evidence that the item is old than new (i.e., the likelihood ratio is greater than 1), the item is judged to be old. This decision rule allows for factors that increase discriminability to simultaneously produce increases in hit rates and decreases in false-alarm rates and maintains the assumption that a single familiarity dimension underlies recognition memory judgments. As applied to the word-frequency effect, single-process theories propose that the study of low- and high-frequency items creates a greater separation of the distributions on which recognition memory decisions are made for low-frequency items relative to high-frequency items. Thus, single-factor theories all propose that there is some characteristic of low-frequency items that makes them more discriminable from one another than high-frequency items when they are studied, such as increased attention (Glanzer et al., 1993), more salient features (Shiffrin & Steyvers, 1997), or less variable representations (McClelland & Chappell, 1998). Single-process theories then assume that when this factor has been com-

Jason Arndt and Lynne M. Reder, Department of Psychology, Carnegie Mellon University.

This research was supported by National Institute of Mental Health Grant 5-T32-MH19983.

We thank Julia Spaniol and J. Neil Bearden for helpful comments on an earlier version of this article, and Aaron Benjamin, William Hockley, and an anonymous reviewer for thoughtful reviews of this article.

Correspondence concerning this article should be addressed to Jason Arndt, who is now at Department of Psychology, Middlebury College, Middlebury, Vermont 05753. E-mail: jarndt@middlebury.edu

bined with a likelihood-ratio decision rule, the observed mirror effect results.

Dual-process explanations of recognition memory (Atkinson & Juola, 1974; Mandler, 1980) maintain that two processes contribute to recognition memory: a fast-acting familiarity process and a slower, more deliberate, recollection process, Consistent with this general proposition, dual-process explanations of the mirror effect (Joordens & Hockley, 2000; Reder et al., 2000) propose that the hit-rate portion of the mirror effect is primarily driven by differences in recollection and the false-alarm portion of the mirror effect is driven by differences in familiarity. Thus, in order to explain the word frequency effect, dual-process theories assume that participants are able to recollect low-frequency items more often than high-frequency items, which produces the hit-rate portion of the mirror effect. Further, dual-process theories assume that high-frequency words are more familiar than low-frequency words, which produces the false-alarm portion of the mirror effect. For example, in their account of the word-frequency mirror effect, Reder et al. (2000) proposed that participants are able to recollect low-frequency words better than high-frequency words because low-frequency words have relatively less contextual competition. Thus, when a low-frequency item is studied, participants have an easier time recollecting that it was experienced in the current experimental context. To explain the false-alarm portion of the mirror effect, Reder et al. proposed that pre-experimental factors, such as a more extensive exposure history for high-frequency words, produce differences in familiarity for low- and high-frequency items, rendering high-frequency items more familiar in general. This heightened level of familiarity for high-frequency items relative to low-frequency items produces the false-alarm differences observed in the word-frequency effect.

ROC Curves and Models of Recognition Memory

One manner in which researchers have evaluated the vitality of models of recognition memory is to examine receiver operating characteristic (ROC) curves. ROC curves illustrate the relationship between hits and false alarms at various levels of response bias. Thus, rather than requiring models of recognition memory to explain performance for a single pair of hit and false-alarm rates, ROC curves require models to account for a range of hit and false-alarm rate pairs, as well as the characteristics of the function relating them to one another. Consequently, empirical ROC curves provide a rigorous test of models of recognition memory, given the models' predictions regarding the nature of the curves. The most common method by which response bias is varied is to request that participants provide confidence ratings for their old–new recognition judgments on a scale with approximately 6–10 points. To construct an ROC curve from confidence rating data, one first plots the hit rate against the false-alarm rate for the most confident old judgment category. Next, one plots the cumulation of the hit rate against the cumulation of the false-alarm rate for the most confident and second most confident old categories. This procedure is repeated until a point has been plotted representing the cumulative hit and false-alarm rates for all but the least confident response category, in which the cumulative hit and false-alarm rates are necessarily 1.0. Thus, a confidence scale with N ratings produces an ROC curve with $N - 1$ points.

Not surprisingly, single- and dual-process models of recognition memory make different predictions about the genesis and form of ROC curves. Specifically, single-process models predict that the ROC curve should be the result of placing decision criteria at various points on a continuous decision axis. These criteria determine the points on the recognition memory decision axis at which participants judge an item to fall in a given response category. Thus, for the most confident response category, any value on the decision axis higher than the most confident response criterion is judged as old with high confidence. Similarly, any value on the decision axis higher than the second most confident response criterion but not higher than the most confident response criterion will be judged as old with the second highest degree of confidence. The proportion of the old items falling above a given criterion corresponds to the hit rate at that level of response bias, and the proportion of new items falling above that same criterion corresponds to the false-alarm rate. Although the shape of the distributions underlying performance differs across models, all of the single-process theories that can produce mirror effects predict that recognition memory ROC curves will be asymmetric about the negative diagonal and convex in probability space. Further, single-process models of the mirror effect predict that ROC curves will be linear when the hit and false-alarm probabilities are transformed into z-coordinates to form a z-ROC curve, a characteristic of many discrimination models based on continuous distributions (Murdock, 1965; Van Zandt, 2000). Finally, the slope of the z-ROC curve produced by recent single-process models is less than 1.0. The basis for this prediction is that the distributions underlying recognition memory performance have different variances, with the variance of the old item distribution being greater than the variance of the new-item distribution (Glanzer et al., 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). These predictions are generally in accord with empirical studies of recognition memory ROCs (Glanzer, Kim, Hilford, & Adams, 1999; Gronlund & Elam, 1994; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992). The general predictions of the single-process models described above are depicted graphically in the top panel of Figure 1.

Dual-process theories propose that ROC curves are produced by the convolution of recollection (a high-threshold process) and familiarity (a continuous, normally distributed process; Yonelinas, 1994).[1] The characterization of recollection as a high-threshold process in dual-process theories identifies it as a memory process with qualitatively different characteristics than familiarity.[2] In

[1] Certainly, other forms of continuously distributed processes could be assumed for familiarity. For simplicity, we discuss familiarity in terms of normally distributed process where the distributions of familiarity have the same variance for old and new items.

[2] The most appropriate characterization of a high-threshold process is somewhat unclear. The traditional view of high-threshold processes is that they are all or none in the sense that either every element of an item's presentation is recollected or none of the elements of an item's presentation is recollected. However, it is probably more correct to characterize the recollect state as the participant being able to recollect the particular detail critical to accurate completion of the memory task at hand and the no-recollect state as the participant being unable to recollect the critical detail for the memory task at hand. Such definitions are more accurate in associating estimates of recollection-based processing with the type of

particular, the characterization of recollection as a high-threshold process indicates that there is a psychological threshold for whether an item is recollected. Items falling above the threshold are recollected, whereas items falling below the threshold are not recollected, with the particular definition of what constitutes "recollection" depending on the memory task at hand (e.g., recollection of an item's presentation or recollection of an item pair being presented together in a particular episode). High-threshold processes produce ROC curves that are linear in probability space and concave in *z*-space, as depicted in the second row of Figure 1 (Swets, 1986). Note that high-threshold processes produce ROC curves with a *y*-intercept that is above zero, which provides an estimate of the probability that an old item is above the threshold (i.e., that it is recollected). Observations below threshold elicit a guess by the participant, leading the ROC curve to be linear, with a slope of $1 - p(\mathrm{R})$, where $p(\mathrm{R})$ is the probability of recollection.

The blend of a high-threshold process and a continuous, normally distributed process that is a characteristic of dual-process models produces an ROC curve that is asymmetric as long as recollection contributes to performance (Yonelinas, 1994). The dual-process explanation of ROC curves assumes that if a test item is recollected, it will be assigned to the most confident old response category in a confidence rating experiment, because recollection is the more certain basis for recognition. Further, some items that are recognized on the basis of familiarity, both old and new, will also be placed in the most certain response category because they are extremely familiar to the participant. The less confident points in the ROC curve will be the result of the continuous, familiarity-driven process only, and will give the ROC curve its convex shape. Thus, dual-process theories produce ROC curves that have a *y*-intercept above zero and possess a convex, but asymmetric, shape in probability space. Further, dual-process models of recognition memory produce *z*-ROC curves that are generally linear with a slight concavity at the lower end of the curve, indicating the contribution of recollection to performance (Yonelinas, 1994). The ROC and *z*-ROC curves predicted by a dual-process model with a normally distributed, equal variance familiarity process are presented in the third row of Figure 1.

Discriminating Between Single- and Dual-Process Models With ROC Curves

As one can see by comparing both the ROC and *z*-ROC curves for single- and dual-process models, the predictions of these two models may not differ greatly unless the contribution of recollection is substantial and the contribution of familiarity is minimal. Indeed, it has proven difficult to discriminate between these two classes of models in studies of item recognition, even when researchers test the models with ROC curves (e.g., Glanzer, Hilford, Kim, & Adams, 1999; Yonelinas, 1999b). However, there are several recent reports in the literature that favor dual-process models of recognition over single-process models.

First, Yonelinas (1997) demonstrated that ROC curves for associative recognition are inconsistent with the predictions of a single-process model but are consistent with the recollection component of a dual-process model. Specifically, Yonelinas (1997) demonstrated that ROCs for associative recognition were largely linear in probability space and curvilinear in *z*-space, results that are in accord with the predictions of a high-threshold model of discrimination. In terms of a dual-process model of recognition, such a result would be taken to indicate that recollection is the dominant memory process contributing to discrimination performance in associative recognition (see Kelley & Wixted, 2001, and Quamme & Yonelinas, 2001, for evidence that recollection does not dominate performance in all associative recognition situations). Second, Yonelinas (1999a) demonstrated that source-discrimination ROCs are also inconsistent with the predictions of a single-process model but are consistent with the predictions of a high-threshold model of discrimination, and therefore the recollection component of dual-process models of recognition memory (see Slotnick, Klein, Dodson, & Shimamura, 2000, for results inconsistent with a high-threshold model of source discrimination). Third, Rotello, Macmillan, and Van Tassel (2000) provided evidence consistent with the contribution of recollection to item memory and inconsistent with the predictions of single-process models.

Rotello et al. (2000) used an item-recognition paradigm in which participants were required to discriminate between studied nouns and plurality-reversed distractor items (e.g., study *frog*, test with *frogs*; Hintzman & Curran, 1994; Hintzman, Curran, & Oppy, 1992). Consistent with the predictions of a high-threshold model of discrimination, the confidence-based ROC curve for plurality discrimination was essentially linear in probability space and concave in *z*-space (Rotello et al., 2000). Further, Rotello et al. demonstrated that the ROC curve relating studied items to plurality-reversed distractor items intercepted the upper *x*-axis at a point less than 1.0. Such a result is consistent with the predictions of a dual-threshold model, in which observations below a low threshold are rejected and observations above a high threshold are accepted (Swets, 1986). Specifically, in a dual-threshold model, a *y*-intercept above zero is a measure of the probability of an observation falling above a high threshold (e.g., the probability an item is recollected as studied), as is the case with a high-threshold model. However, a dual-threshold model proposes that participants systematically reject some observations that fall below a low threshold. Thus, rather than the ROC curve intercepting the upper *x*-axis at 1.0, as is the case for a high-threshold model, the systematic rejection of some observations will produce an ROC curve that intercepts the upper *x*-axis at a point less than 1.0 (referred to as the *upper x-intercept* below). Further, the deviation of the upper *x*-intercept from 1.0 is an index of the probability of an observation falling below the lower threshold. The ROC and *z*-ROC curves predicted by a dual-threshold model are presented in the bottom row of Figure 1.

In terms of the recollection process, this result is consistent with the presence of a *recall-to-reject* strategy (Clark & Gronlund, 1996; Hintzman & Curran, 1994; Rotello et al., 2000). Such a strategy is based on the notion that participants utilize their ability to recall studied items to reject similar distractors. Thus, for

discrimination required by a given memory task. The important point for the present analysis is that high-threshold memory processes are qualitatively different than the memory process embodied in most continuous-distribution models, such as the standard Signal Detection Theory (SDT) model with Gaussian distributions, and therefore produce ROCs that have qualitatively different characteristics than extant single-process models.

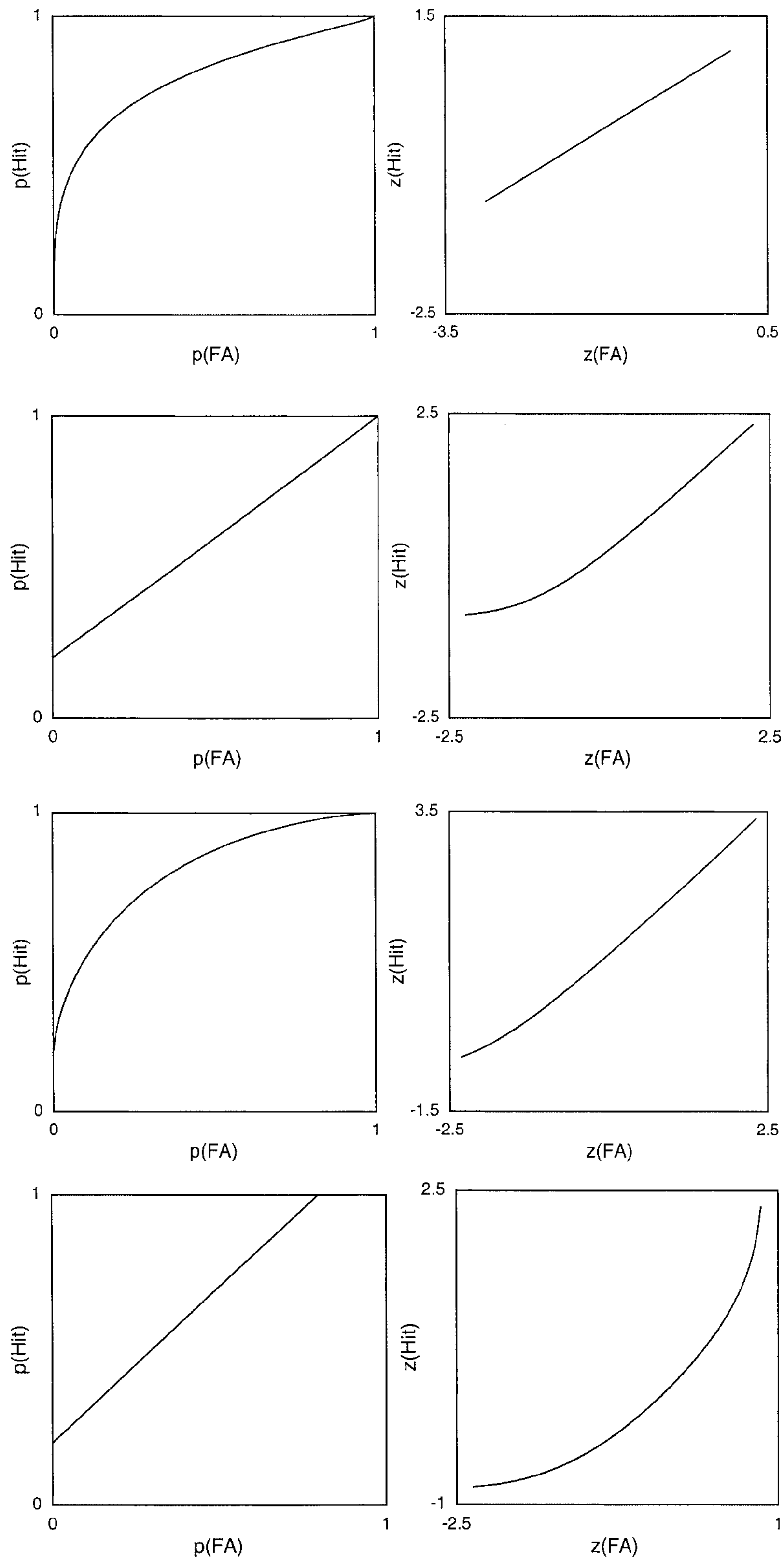

Figure 1. Predicted receiver operating characteristic (ROC; left column) *z*-transformed ROC (*z*-ROC; right column) curves for single-process models (top row), high threshold models (second row), dual-process models (third row), and dual threshold models (bottom row).

example, if a participant was shown the word *books* at study and was presented with the word *book* at test, she or he may be able to reject the test item as unstudied on the basis of the ability to recall that *books* was studied rather than *book* and therefore would reject the item with high confidence. Additionally, and critical to our use of this paradigm to study the word-frequency effect, the greater the contribution of recollection to performance, the greater the *y*-intercept. Similarly, the more often a recall-to-reject strategy is utilized, the more the upper *x*-intercept will deviate from 1.0.

The Present Experiments

In these two experiments, we test the explanation of recent dual-process models of the word-frequency effect in recognition memory (Joordens & Hockley, 2000; Reder et al., 2000). In the first experiment, we evaluate these dual-process models by using the paradigm of Rotello et al. (2000) and manipulating word frequency. Thus, we construct a situation in which discrimination between studied items and some lure items should be extremely difficult, likely requiring recollection. The construction of this condition is designed to test two predictions of dual-process theories of the word-frequency effect. The first prediction is that recollection differences between low- and high-frequency items produce the observed differences in hit rates. If this prediction is correct, low-frequency items should show greater recollection than high-frequency items on the basis of the characteristics of the ROC curves relating recognition of studied items to erroneous recognition of plurality reversed distractors. Specifically, the ROCs for low-frequency words should have higher *y*-intercepts and lower upper *x*-intercepts relative to ROCs for high-frequency words, indicating that participants were able to utilize recollection more for low- than high-frequency words. The second prediction is that the false-alarm portion of the word-frequency effect arises from differences in familiarity. If this prediction is correct, low- and high-frequency items would be expected to show comparable false-alarm rates when they are plurality-reversed lure items. Specifically, because the rejection of plurality-reversed lure items should be primarily driven by recollection, differences in familiarity should not contribute to performance, producing an equivocation of the false-alarm rates for low- and high-frequency items.[3]

The second experiment verifies that our stimulus materials produce the traditional word-frequency mirror effect in terms of hits and false alarms when participants are required to discriminate only between studied items and entirely unstudied new items. Additionally, in the second experiment, we tested whether our stimulus materials show the same characteristics of the confidence-based *z*-ROC curves that have been observed by other researchers. Specifically, the intercept of the *z*-ROC has been shown to be higher for low- than for high-frequency items, whereas the slope of the *z*-ROC has been shown to be lower for low- than for high-frequency items (Glanzer et al., 1999; Ratcliff et al., 1994).

Experiment 1

In this experiment, participants were presented with nouns in either their singular or plural form at study. At test, three different types of items were presented: studied items, plurality-reversed distractor items, and entirely novel distractor items. The use of singular nouns and their plural forms afforded us the opportunity to most effectively study the contributions of recollection to recognition memory, because effective discrimination between studied items and plurality-reversed distractor items should require the use of recollection (Hintzman & Curran, 1994; Hintzman et al., 1992; Rotello et al., 2000).

Method

Participants. Thirty-five students at Carnegie Mellon University participated in order to fulfill a research appreciation requirement.

Materials and design. The stimulus materials were 180 low-frequency and 180 high-frequency nouns and their plural forms (Kučera & Francis, 1967). Item pairs were selected such that they would be as semantically and orthographically similar as possible. To this end, stimulus pairs were required to meet three criteria. First, the plural version of each item could be created by adding *s*. Second, the dominant meaning of each item was the same in its singular and plural form. Third, the singular and plural form fell within the same frequency category. Thus, pairs were rejected if they consisted of a low-frequency singular form and a high-frequency plural form or vice versa. Low-frequency items occurred fewer than 4 times per million words, and high-frequency items occurred greater than 24 times per million words. Singular low-frequency items had a mean frequency of 1.71, plural low-frequency items had a mean frequency of 1.66, singular high-frequency items had a mean frequency of 154.22, and plural high-frequency items had a mean frequency of 78.34.

The design formed a 2 × 3 factorial, with both word frequency (high vs. low) and test item type (old vs. similar vs. new) manipulated within participants. Stimulus items were divided into three lists of words, each with 60 low-frequency and 60 high-frequency singular–plural pairs. Each of the three lists of words was further divided into three sets of item pairs to serve in the three test item type conditions, with 20 low-frequency and 20 high-frequency pairs assigned to each condition (old, similar, or new). Assignment of item pairs to one of the three stimulus lists and to one of the three experimental conditions was determined randomly for each participant.

Items assigned to the old and similar conditions within each study list were presented to participants in a study list, whereas items assigned to the new condition were reserved for presentation in the test list only. For items in all three conditions, half were the singular form and half were the plural form of that item pair. Thus, study lists were composed of 20 low-frequency items in their singular form, 20 low-frequency items in their plural form, 20 high-frequency items in their singular form, and 20 high-frequency items in their plural form. Additionally, two primacy and two recency buffers of medium frequency were added to the study list, yielding a list length of 84 items. At test, participants were presented with items identical to their studied form (old items), items similar to ones which had been studied but with the opposite plurality (similar items), and items which had not been studied either in part or in whole (new items). Memory for buffer items was not tested, yielding a test list length of 120 items (40 old items, 40 similar items, and 40 new items), with half of the items in each condition being low frequency and half being high frequency. Further, half of the items in each of the six cells of the design (two levels of word frequency crossed with three levels of test item type) were singular and half were plural. Assignment of items to experimental conditions and serial position in both the study and test lists was determined randomly for each participant.

Procedure. Participants completed three study test cycles in which all aspects of the procedure were essentially the same. Prior to each study list, participants were instructed that they would be shown a list of words sequentially on the computer screen and that their task was to remember

[3] We thank Aaron Benjamin for making this observation.

the words for a later memory test. The study items were then shown serially in the center of a computer screen for 2 s each. Immediately following the presentation of each study list, participants were presented with the recognition memory test instructions. Participants were asked to provide confidence ratings of whether each item had been studied, on a six-point scale (1 = *sure old*; 6 = *sure new*). Participants were additionally informed that there would be three types of items on the memory test: study items, which were to be called "old"; entirely unstudied items, which were to be called "new"; and similar items, which were also to be called "new". They were provided with an example of a similar test item and informed that only one form of each item would have been presented to them on the study list, and only one form would appear on the test list. Consequently, if they remembered that a particular test item's opposite plurality form had been studied, they could be certain it was a new item, and reject it with high confidence (i.e., they could assign a "sure new" response to it). Participants progressed through the test list at their own pace. Following the completion of the first test list, a new study list was presented. Prior to the presentation of the second and third study lists, participants were instructed that their memory for the previous study lists would not be tested again and that only the items on the present study list would be tested on the upcoming memory test. The experiment concluded when participants had completed three study–test cycles.

Results and Discussion

Hit and false-alarm analyses. The mean hit rate for old items and the mean false-alarm rates for similar and new items are presented in Figure 2. As is evident, old-item hits and new-item false-alarms showed the standard word-frequency effect, with hits for low-frequency items being greater than hits for high-frequency items, whereas false alarms for low-frequency items were lower than false alarms for high-frequency items, both $t(34) > 5.02$. However, false alarms to similar items did not reliably differ as a function of word frequency, $t(34) = 1.10$, $p > .25$. As noted above, this latter result is expected by dual-process models' account of the word-frequency effect. Specifically, such models propose that false-alarm differences between low- and high-frequency items arise because of differences in familiarity. Therefore, in a situation in which discrimination is accomplished primarily or entirely based on recollection, such models would not expect a difference in the false-alarm rates between items of different word frequency, a prediction consistent with the result observed for false alarms to similar items.

Given the theoretical importance of this latter result, we sought to ensure that the lack of a significant difference between false alarms to low- and high-frequency similar items was not simply a product of low statistical power. We therefore conducted a power analysis on the comparison between false alarms to low- and high-frequency similar items. We used the effect size found for the false-alarm rates to new items as our benchmark for power computations ($f = 1.31$). The power to detect a false-alarm-rate difference one half the size of that found for new items in this design was in excess of .98, and the power to detect a false-alarm-rate difference one third the size of that found for new items was in excess of .81, both of which are traditionally classified as high levels of power (Cohen, 1988). Thus, we believe there was sufficient power to detect a reasonably sized false-alarm-rate difference between low- and high-frequency similar items.

Form of the ROCs, z*-ROCs, and appropriate discrimination models.* Following Rotello et al. (2000), we constructed two types of ROC curves for these data. The first plots hits against false

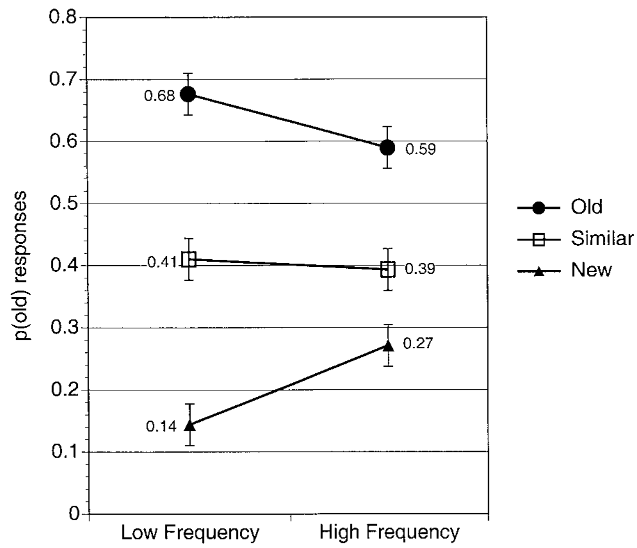

Figure 2. Hits to old items and false alarms to similar and new lure items in Experiment 1 as a function of word frequency. Error bars depict 95% confidence intervals.

alarms to entirely new items (referred to as an *old–new ROC* below). The second plots hits against false alarms to similar items (referred to as an *old–similar ROC* below). Further, *z*-ROCs were constructed by converting each participants' cumulative hit and false-alarm rate into a *z* score and plotting the function relating the *z*-transformation of the hit rate to the *z*-transformation of the false-alarm rate. Recall that continuously distributed processes show a linear relationship between hits and false alarms in *z*-space whereas threshold processes show a concave relationship between hits and false alarms in *z*-space.

Old–new and old–similar *z*-ROCs are presented in Figure 3 as a function of word frequency. Note that the old–new *z*-ROCs are well approximated by a linear fit, whereas the old–similar *z*-ROCs show a marked concavity. To quantitatively evaluate the linearity of the functions of the old–new and old–similar *z*-ROCs, we regressed hits on false alarms for each participant's old–new and old–similar *z*-ROCs and included both linear and quadratic terms in the regression equation.[4] If a linear trend is sufficient to describe the relationship between hits and false alarms, the expected value of the quadratic terms is zero. However, if the ROC curve shows a concave pattern, the expected value of the quadratic terms is positive. Old–new *z*-ROCs for high- and low-frequency items failed to show reliable evidence of curvature—mean quadratic = 0.01 for high-frequency items and −0.02 for low-frequency items; both $t(33) < 0.72$—indicating that the *z*-ROCs were accurately described by a linear trend. In contrast, old–similar *z*-ROCs showed evidence of a concave shape for both high- and low-frequency items—mean quadratic = 0.18 and 0.34 for high- and low-frequency items, respectively; smallest

[4] One participant categorized all new low-frequency items as "sure new" making it impossible to construct an individual old–new ROC curve. Thus, analyses of the characteristics of old–new *z*-ROCs were based on 34 of the 35 participants in this experiment.

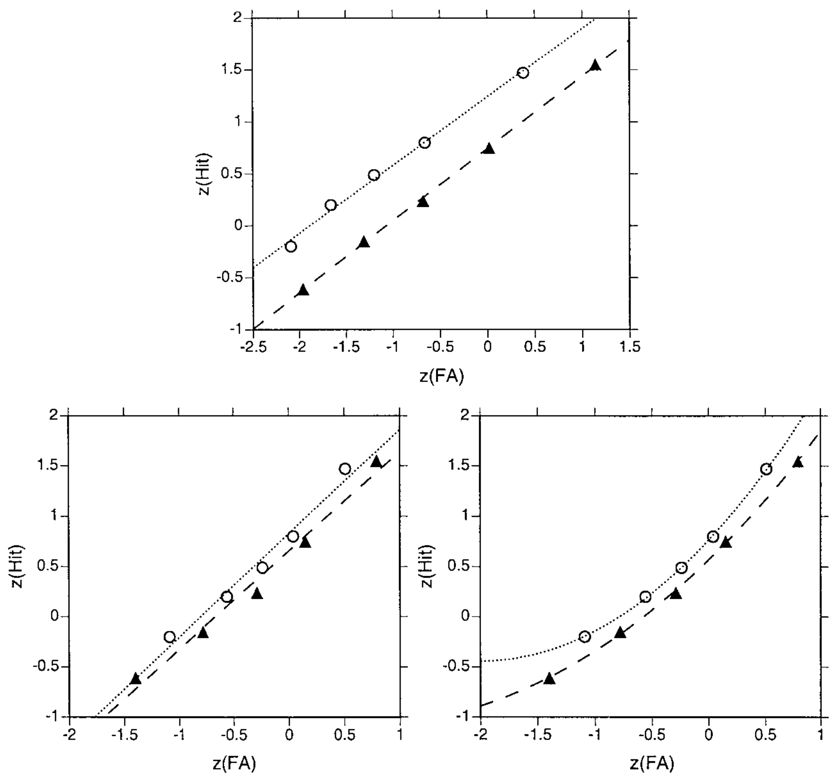

Figure 3. *z*-transformed receiver operating characteristics (*z*-ROCs) from Experiment 1 as a function of word frequency. Triangles represent performance for high-frequency items, open circles represent performance for low-frequency items. Functions for low-frequency items are dotted and functions for high-frequency items are dashed. The top panel depicts the *z*-ROCs for old–new discrimination, with the best-fitting linear trend. The bottom two panels depict the *z*-ROCs for old–similar discrimination. The left panel depicts the best-fitting linear trend and the right panel depicts the best-fitting regression model with quadratic components.

$t(34) = 4.19$. Thus, old–new discrimination performance is consistent with the predictions of single-process models of recognition memory. However, old–similar discrimination performance is inconsistent with the predictions of single-process models of recognition memory but is consistent with the predictions of threshold models of discrimination.

Further support for this conclusion comes from informal analyses of the old–new and old–similar ROCs, which are presented in Figure 4. Comparison of these figures reveals that whereas the old–new ROCs showed the concave downward pattern typically observed in recognition memory experiments, the old–similar ROCs showed a considerably more linear relationship between hits and false alarms. In an effort to illustrate these differences, we plotted the best-fitting ROC curves produced by the Rockit maximum-likelihood estimation algorithm (Metz, 1998) in Figure 4. Rockit assumes that normal distributions underlie performance in a discrimination task, and therefore the algorithm utilized by Rockit will necessarily produce a curve that has the characteristics of a model of discrimination with a continuous, normally distributed process.[5] Note that the best-fitting ROC curves produced by Rockit describe the old–new data well, whereas the old–similar ROCs appear to be more linear than would be expected on the basis of the best-fitting Rockit solution.

In summary, the old–new ROCs and *z*-ROCs are consistent with the predictions that current single-process models make for old–new recognition (e.g., Glanzer et al., 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Support for this conclusion comes from the fact that the old–new *z*-ROCs were accurately approximated by a linear function. Further, the best-fitting ROC curves produced by Rockit approximated old–new recognition data quite well. However, the old–similar ROCs and *z*-ROCs are at variance with the predictions of single-process models. Specifically, *z*-transformation of the hit and false-alarm rates for these ROCs produced a concave relationship, whereas a model based on a normally distributed process should produce a linear *z*-ROC. Further, the old–similar ROCs were not accurately described by a best-fitting ROC solution from Rockit, showing more linearity than would be expected if the distributions underlying performance were normal. Although single-process models of recognition memory do not necessarily assume that the familiarity distributions underlying performance are normal in shape, they are all

[5] In actuality, Rockit and other maximum-likelihood estimation algorithms assess the characteristics of ROC curves by assuming that logistic distributions underlie performance. The logistic distribution is a mathematically simple approximation to the normal distribution, and models assuming underlying logistic distributions produce ROC and *z*-ROC curves that are essentially indistinguishable from models assuming underlying normal distributions.

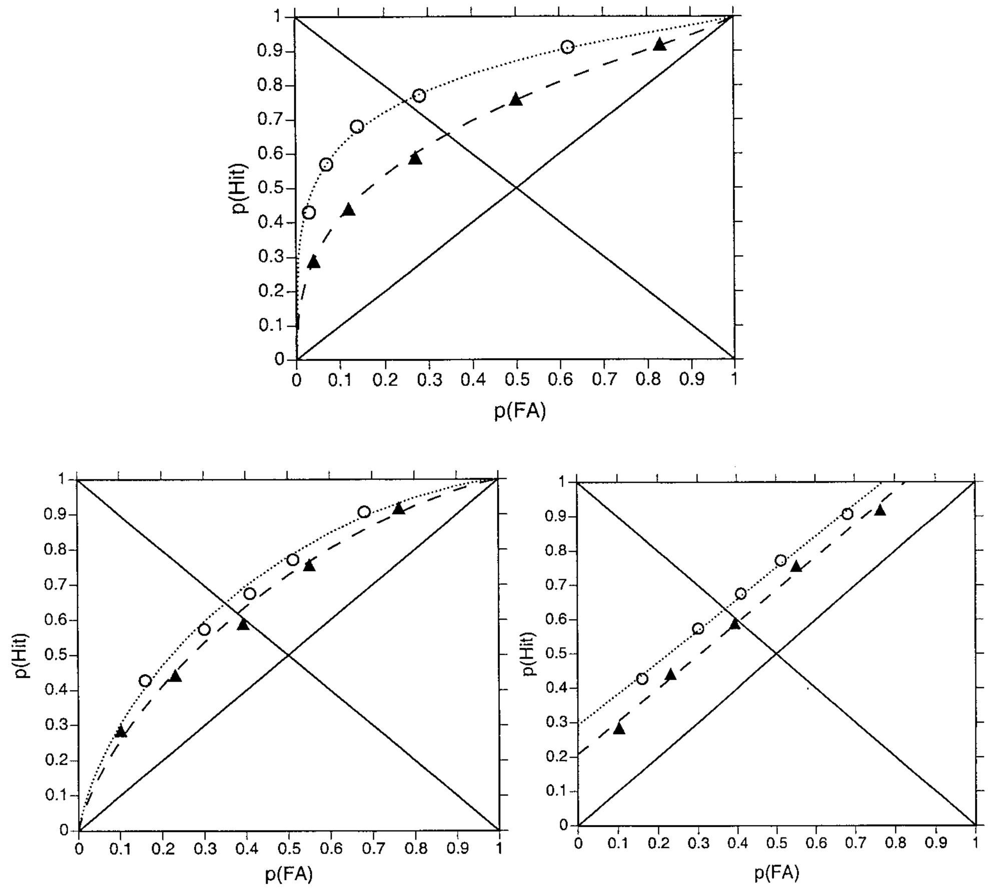

Figure 4. Receiver operating characteristics (ROCs) from Experiment 1 as a function of word frequency. Triangles represent performance for high-frequency items, open circles represent performance for low-frequency items. Functions for low-frequency items are dotted and functions for high-frequency items are dashed. The top panel depicts the ROCs for old–new discrimination, with the best-fitting ROC function generated by the Rockit maximum-likelihood estimation algorithm. The bottom two panels depict the ROCs for old–similar discrimination. The left panel depicts the best-fitting ROCs generated by the Rockit maximum-likelihood estimation algorithm and the right panel depicts linear regression fits.

constrained to produce ROC curves that are of the type we observed for old–new recognition (Glanzer et al., 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). That is, all current single-process models are constrained to produce ROC curves that are convex in probability space and linear in *z* space. Thus, none of the extant single-process models of recognition memory are capable of producing the type of ROC curve we observed for discrimination between old items and similar distractors with a familiarity process alone.

Although the old–similar ROCs are inconsistent with the predictions of extant single-process models of recognition memory, they are consistent with the predictions of a dual-threshold model of discrimination. Recall that a dual-threshold model is a variant of a high-threshold model of discrimination in which some distractor items are systematically rejected. Thus, dual-threshold models predict (a) a linear relationship between hits and false alarms and (b) that the function relating hits to false alarms intercepts the upper *x*-axis at a point below 1.0. Further, a hallmark prediction of the dual-threshold model is a concave *z*-ROC curve, which is consistent with the result observed here in the old–similar *z*-ROC curves for both low- and high-frequency words. Consequently, it is reasonable to propose that a dual-threshold model accurately characterizes the memory process that mediates discrimination between old items and similar distractor items.

The memory process we assume to mediate discrimination between old items and similar distractor items is a variant of the recollection component of dual-process models of recognition memory (Atkinson & Juola, 1974; Mandler, 1980) in which recollection can be used both to affirm that a study item is old and to reject similar distractor items (Clark & Gronlund, 1996; Hintzman & Curran, 1994; Rotello et al., 2000). We proceed next to analyze the estimates of recollection for low-frequency and high-frequency items derived from the characteristics of a dual-threshold model. Then, we compare those results with the predictions of recent dual-process explanations of the word-frequency effect (Joordens & Hockley, 2000; Reder et al., 2000).

Analyses of dual-threshold model parameters. In the previous section, we analyzed the characteristics of the old–similar ROCs and found evidence consistent with the predictions of a dual-threshold model of performance. Here, we analyze differences between the old–similar ROCs for high- and low-frequency items. Recall that the dual-process explanation of the word-frequency effect proposes that the hit-rate portion of the word-frequency effect is produced by differences in recollection between low- and high-frequency items. If the dual-process explanation of the word-frequency effect is correct, we should find evidence of greater recollection for low-frequency items in the old–similar ROCs because we assume performance in this condition to reflect the effects of recollection on recognition memory.

If a dual-threshold model alone is sufficient to describe discrimination between old items and plurality-reversed lures, the *y*-intercept and upper *x*-intercept of the old–similar ROC should provide estimates of the amount of recollection that was available to participants for old low- and high-frequency items. Consistent with the dual-process model's explanation of the word-frequency effect for hits, the *y*-intercept of the old–similar ROC for low-frequency items was higher than for high-frequency items (.29 vs. .21). Similarly, the upper *x*-intercept was lower for low- than for high-frequency items (.77 vs. .83), indicating that participants were able to use recollection to reject similar distractor items more often for low-frequency items. Thus, the estimates of recollection based on the *y*-intercept would indicate that participants in this experiment were able to use recollection to accept study items about 29% of the time for low-frequency words and about 21% of the time for high-frequency words. Similarly, the estimates of recollection based on the upper *x*-intercept indicate that participants were able to use recollection to reject distractors about 23% of the time for low-frequency items and about 17% of the time to reject high-frequency distractors.

We verified that these results were reliable in two different ways. First, we used linear regression to predict the *y*-intercept and upper *x*-intercept of the old–similar ROC curve for each participant. This analysis produced reliable differences in both the *y*-intercepts, .30 vs. .22; $t(34) = 3.49$, and upper *x*-intercepts for low- and high-frequency items, .78 vs. .84; $t(34) = 2.96$, with both measures indicating greater recollection for low-frequency items. Second, we analyzed a measure of discriminability that is appropriate for a dual-threshold model, $H'c$: $H'c = P(\text{hit}) - P(\text{false alarm})$ (Swets, 1986). The dual-threshold model for which this measure of discriminability is appropriate assumes that $H'c$ is invariant across levels of response bias. Therefore, the estimate of recollection that is derived from this correction procedure should be constant across the different confidence ratings (i.e., the slope relating hits to false alarms should be 1.0). We assessed these predictions with a 2 (low vs. high frequency) × 5 (confidence category) analysis of variance ANOVA using $H'c$ as the dependent measure. This analysis revealed a main effect of word frequency, $F(1, 34) = 18.198$, $MSE = 0.223$; and a main effect of confidence category, $F(4, 136) = 8.419$, MSE = 0.0034, but no interaction ($F < 1$). The main effect of word frequency indicates that $H'c$ was higher for low- than for high-frequency items, and the lack of an interaction indicates that the difference in $H'c$ between low- and high-frequency items was approximately constant across all five levels of confidence. These two results are consistent with our analysis of the *y*-intercept and the upper *x*-intercept for the old–similar ROCs and are consistent with the conclusion that recollection was greater for low- than for high-frequency items. Further, paired comparisons between $H'c$ for low-frequency items versus high-frequency items revealed that $H'c$ was reliably higher for low- than for high-frequency items in all five confidence categories, smallest $t(34) = 2.73$, $p < .01$.

However, the main effect of confidence category reveals that $H'c$ differed across levels of response bias, in contrast to the prediction of the dual-threshold model for which $H'c$ is an appropriate measure of discrimination. There are two potential reasons for this. First, it could be the case that recollection is greater when accepting studied items than when rejecting similar lure items, leading the old–similar ROC curve to have a slope less than 1.0. This proposal seems reasonable, given that one may expect plurality reversed distractor items to be slightly poorer retrieval cues than old items because old items replicate the orthography and semantics of study items exactly whereas similar items deviate slightly in terms of both orthography and semantics from study items. A second potential reason for this difference is that discrimination between old items and similar items may not be entirely based on recollection. That is, old items could have marginally greater levels of familiarity than plurality-reversed distractors because the test probe matches slightly better to studied items than to similar distractor items. This small contribution of familiarity could influence the hit rate more than the false-alarm rate to similar items, leading to the slight, but reliable, effect of response category on $H'c$. Regardless, the characteristics of the old–similar ROCs are largely in accord with the predictions of the dual-threshold model that we assume to describe discrimination between old items and similar items. Further, our three potential measures of recollection, the *y*-intercept, the upper *x*-intercept, and $H'c$ are all consistent with one another in describing the differences between the old–similar ROC curves for low- and high-frequency items. Thus, all of these measures are consistent with the same interpretation of participants' discrimination between studied items and similar distractors—specifically, that participants were able to recollect low-frequency items more often than high-frequency items.

Dual-process model analyses. Finally, we analyzed the old–new ROCs in terms of the dual-process model of Yonelinas (1994, 1999a, 1999b) to provide model-based estimates of recollection and familiarity. This analysis served two purposes. First, the estimates of recollection derived from this model should converge with the dual-threshold model analyses presented above, indicating that recollection was greater for low- than for high-frequency items. Second, this estimation procedure provides information that is not available from the above analyses. Specifically, although the results of this experiment strongly indicate that low- and high-frequency items differ in terms of recollection, it is an open question as to whether low- and high-frequency items also differ in terms of incremental familiarity resulting from study.

To derive estimates of recollection and familiarity, we followed the model fitting procedures of Yonelinas (1999b). In particular, we constructed a set of equations describing performance at each of the five points on the old–new ROC curves in these data. The model's equation describing hit rates is

$$R + (1 - R)\phi\left(\frac{d'}{2} - c_i\right), \qquad (1)$$

where R is the probability of recollection, d' is the standard distance between the old and new familiarity distributions in a Gaussian equal-variance signal-detection model, and c_i is the standardized measure of criterion placement for each point on the ROC curve. The model's equation for false-alarm rates is

$$\phi\left(\frac{-d'}{2} - c_i\right), \qquad (2)$$

where d' and c_i are the same as in Equation 1. Thus, five equations for each hit rate and five equations for each false-alarm rate were constructed, one for the hit and false-alarm rate at each point on the ROC, with the only difference across equations being the placement of c_i. The model was fit by minimizing the sum of squared deviations between each participants' performance and the model with Microsoft Excel's Solver (see Dodson, Prinzmetal & Shimamura, 1998, for a comparison of the derivation of model parameters with Excel's Solver (2001) and maximum likelihood estimation).

The results of this analysis indicated that both recollection (R; .43 vs. .26) and familiarity (d'; 0.95 vs. 0.50) were found to be greater for low- than for high-frequency items, smallest $t(34) = 5.77$. Thus, the analysis of the estimates of recollection derived from the old–new ROCs converges with the conclusions based on the analysis of old–similar ROCs, again indicating that recollection was greater for low- than for high-frequency items. Further, on the basis of the assumptions of this dual-process model, the increment in familiarity resulting from the presentation of study items was also found to be greater for low- than for high-frequency items, a conclusion consistent with other measurement techniques (e.g., process dissociation and remember–know judgments; Yonelinas, 2002).

Experiment 2

The goal of this experiment was to verify that our stimulus materials show similar characteristics of other manipulations of word frequency reported in the literature. First, these materials should show a mirror effect in old–new recognition. Thus, the hit rate for low-frequency items should be higher than the hit rate for high-frequency items and the false-alarm rate for low-frequency items should be lower than the false-alarm rate for high-frequency items. Second, we would expect to replicate the previous findings reported in ROC experiments that manipulated word frequency: specifically, that z-ROC curves for low-frequency items show a lower slope and a higher intercept relative to high-frequency items (Glanzer et al., 1999; Ratcliff et al., 1994). Third, we again fit Yonelinas's (1994, 1999a, 1999b) dual-process model to these data, with the expectation that the parameter estimates derived from the model fitting procedure would be qualitatively similar to those derived from Experiment 1.

Method

Participants. Twenty students at Carnegie Mellon University participated in order to fulfill a research appreciation requirement.

Materials and design. The stimulus materials were the same as those used in Experiment 1. The design formed a 2 × 2 factorial, with word frequency (high vs. low) and test item type (old vs. new) manipulated within subjects. Ninety low-frequency items and 90 high-frequency items were randomly selected to serve as study items, with the remaining 90 items of each stimulus class chosen to serve as new items on the recognition memory test. This yielded a study list length of 180 items and a test list length of 360 items. Half of the old and new items of each stimulus class were singular and half of the items of each stimulus class were plural. Additionally, only one form of each of the word pairs was presented in the study and test list. Therefore, if the singular version of a word pair had been presented as a study item, the plural form would not be presented in either the study or test list. Assignment of items to be studied or unstudied and assignment to serial position in both the study and test lists was determined randomly for each participant.

Procedure. At the beginning of each experimental session, participants were instructed that they would be shown a list of words sequentially on the computer screen and that their task was to remember the words for a later memory test. The study items were then shown serially in the center of a computer screen for 2 s each. Immediately following the presentation of each study list, participants were presented with the recognition memory test instructions. Participants were instructed that their task was to judge whether items had been studied on the list of words they had just been presented with. Further, participants were asked to provide confidence ratings of whether each item had been studied, on a six-point scale (1 = *sure old*; 6 = *sure new*). Participants were then allowed to progress through the test list at their own pace. The experimental session concluded when participants had provided judgments for all of the test items.

Results and Discussion

The results of this experiment are relatively straightforward. These data show evidence of a mirror effect, where hit rates were higher (.63 vs. .58), $t(19) = 2.32$, and false alarm rates were lower (.20 vs. .37), $t(19) = 6.56$, for low-frequency items relative to high-frequency items. The slopes and intercepts of the confidence-based z-ROC curves were estimated for each participant separately for low- and high-frequency items using both linear regression and maximum-likelihood estimation (Dorfman & Alf, 1969; Ogilvie & Creelman, 1968).[6] Analysis of the slopes and intercepts were then compared by using paired-samples t tests to contrast the slopes and intercepts of the high- and low-frequency z-ROCs. The conclusions reached based on linear regression analyses and maximum-likelihood estimation were identical, thus we present only the slope and intercept parameters derived from maximum-likelihood estimation. The z-ROCs for low-frequency items had a lower slope (.64 vs. .77), $t(18) = 3.47$, and a higher intercept (1.01 vs. .64), $t(18) = 4.21$, than the z-ROCs for high-frequency items, replicating the pattern observed in previous studies of recognition memory (Glanzer et al., 1999; Ratcliff et al., 1994). Thus, our stimulus materials appear to show the same characteristics that previous manipulations of word frequency have shown in the literature.

As with the first experiment, we fit the dual-process model of Yonelinas (1994, 1999a, 1999b) to the old–new ROCs for this experiment. The results of this analysis converge with the results of Experiment 1. Specifically, we found that the estimates of both recollection (R; .45 vs. .30) and familiarity (d'; 0.52 vs. 0.20) provided by this model were greater for low- than for high-frequency items, smallest $t(18) > 3.74$. Thus, consistent with the conclusions from Experiment 1, recollection was found to be

[6] One participant categorized all new low-frequency items as "sure new" making it impossible to construct an individual z-ROC curve. Thus, analyses of slopes and intercepts were based on estimates from 19 of the 20 participants in this experiment.

greater for low- than for high-frequency items. Further, the increment in familiarity following study was also found to be greater for low- than for high-frequency items, again replicating both the pattern found in Experiment 1, as well as that reported by Yonelinas (2002).

General Discussion

The results of these two experiments address three critical issues we wish to emphasize. First, the characteristics of the old–similar ROCs are consistent with the view that recollection is a high-threshold process, as is posited in many dual-process models of recognition memory (e.g., Yonelinas, 1994). Second, these results provide support for recent theoretical interpretations of the word-frequency effect, derived from dual-process models of recognition memory (Joordens & Hockley, 2000; Reder et al., 2000). Third, these results pose a critical challenge to models of recognition memory (Benjamin et al., 1998; Gillund & Shiffrin, 1984; Glanzer et al., 1993; Hintzman, 1988; McClelland & Chappell, 1998; Murdock, 1982; Shiffrin & Steyvers, 1997) in which a unitary process is proposed to underlie recognition memory performance. Each of these points is discussed in turn.

The underlying assumption in most dual-process theories is that two memory processes with qualitatively different characteristics contribute to recognition memory performance (Atkinson & Juola, 1974; Jacoby, 1991; Joordens & Hockley, 2000; Mandler, 1980; Reder et al., 2000). "Familiarity" is often characterized as a process based on a continuous measure of memory strength, in which some items and item classes are more familiar than others. However, "recollection" is often characterized as a process that is considerably more certain than familiarity (Jacoby, 1991; Mandler, 1980; Yonelinas, 1994) and requires search for and retrieval of an encoding episode (e.g., Reder et al., 2000). A plausible model of the recollection process is a high-threshold model. Such a theoretical model predicts a linear ROC curve in probability space and a concave ROC curve in *z*-space (Green & Swets, 1966; Swets, 1986). In the present experiment, the observed old–similar ROC curves were consistent with the ROCs predicted by a dual-threshold model, a variant of a high-threshold model. In the context of recognition memory, a plausible dual-threshold memory process is one in which recollection can be utilized both to affirm an item was studied and to reject similar items that were not studied (Clark & Gronlund, 1996; Hintzman & Curran, 1994; Rotello et al., 2000).

Recent logical extensions of the dual-process model of recognition memory have proposed that the mirror effect in recognition memory can be understood in terms of the effects of recollection on hits and the effects of familiarity on false alarms (Joordens & Hockley, 2000; Reder et al., 2000). Thus, these extensions of dual-process theory propose that accurate recognition is significantly affected by the influences of recollection. In terms of the word-frequency effect, these extensions propose that recollection should be greater for low-frequency items than for high-frequency items. In the present experiments, we provided evidence consistent with this hypothesis in terms of the characteristics of the old–similar ROC curves. The analysis of the characteristics of these curves indicated greater ability to recollect low-frequency items than high-frequency items, precisely the result predicted by dual-process explanations of the word-frequency effect (Joordens & Hockley, 2000; Reder et al., 2000). Further, these dual-process models propose that familiarity differences lead to the false-alarm portion of the word-frequency effect. Analysis of the false-alarm rates for similar lure items in our first experiment provided support for this prediction. Specifically, those results indicated that false alarms to similar lure items did not vary as a function of word frequency. Given the assumption that the rejection of such lures is accomplished primarily via recollective processes, an assumption supported by the old–similar ROC analyses, this result is consistent with dual-process explanations of the word-frequency effect. In particular, when familiarity is not utilized for discrimination, false-alarm rates would not be expected to vary as a function of word frequency, exactly the result observed in our first experiment.

Two of the results from these experiments pose difficulty for single-process accounts of the word-frequency mirror effect. First, single-process explanations of the mirror effect (Glanzer et al., 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) all predict ROC curves that are clearly inconsistent with the old–similar ROC curves observed in our first experiment. Specifically, extant single-process models explain the word-frequency mirror effect by assuming that recognition memory is based on a single continuous familiarity dimension. These models are all constrained to produce ROC curves that are (a) convex in probability space, (b) typically asymmetric about the negative diagonal, and (c) linear in *z*-space. Thus, these single-process models are unable to account for item-recognition performance when the discrimination between studied and unstudied items is difficult, as is the case for discrimination between old items and similar lure items. Second, the manner in which extant single-process models have accounted for the occurrence of mirror effects is to induce a dependency between hits and false alarms. Thus, factors that increase hits also decrease false alarms in these models. Consequently, the false-alarm rates to similar lure items are also problematic for these models, because those false alarms were found to vary independent of hit rates as a function of word frequency.

An objection one may have to the present results is that in order to produce evidence that is clearly at variance with single-process models of recognition memory but that is consistent with dual-process models, we were required to construct a situation in which familiarity-based discrimination would be largely unsuccessful. This argument ignores the significant theoretical contribution that dual-process explanations offer for these data. Specifically, dual-process explanations of the word-frequency effect provide a prediction that the hit-rate advantage for low-frequency words should be due to differences in recollection. Thus, if one is able to construct a situation in which recollection is the dominant basis of discrimination, one should uncover evidence consistent with greater levels of recollection for low- than for high-frequency words. In the first experiment described above, such evidence was found in terms of the characteristics of the old–similar ROCs for low- and high-frequency items, as well as the false-alarm rates for similar lure items. Further, in a situation where recollection is the primary determinant of performance, dual-process explanations of the word-frequency effect would expect that false alarms would not vary as a function of word frequency, a result that was also confirmed in the first experiment reported here.

In concluding, we wish to emphasize the constraints that these results place on theories of recognition memory. First, comprehensive theories of recognition memory must be able to simultaneously account for both the characteristics of performance on old–new discrimination and old–similar discrimination observed in these experiments. Specifically, theories of recognition memory must be able to produce both (a) the ROCs typically observed for discrimination between studied items and entirely novel distractors and (b) the ROCs observed for discrimination between studied items and similar distractors. Further, theories of recognition memory must account for not only the traditional false-alarm rate differences between low- and high-frequency items but also why the false-alarm rates are comparable for plurality-reversed distractors as a function of word frequency. Second, theories of recognition memory must also account for the manner in which these two different ROC curves vary across experimental conditions. On the basis of these data, theories of recognition memory must explain why discrimination between studied and unstudied items was better for low- than for high-frequency items regardless of whether the discrimination was relatively easy (old–new recognition) or difficult (old–similar recognition). Taking into account the characteristics of old–new and old–similar discrimination, as well as the manner in which such discrimination performance varied as a function of word frequency, these data favor a dual-process interpretation (e.g., Joordens & Hockley, 2000; Mandler, 1980; Reder et al., 2000; Yonelinas, 1994).

References

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 1. Learning, memory, and thinking* (pp. 243–293). New York: Freeman.

Benjamin, A. S., Bjork, R. A., & Hirshman, E. (1998). Predicting the future and reconstructing the past: A Bayesian characterization of the utility of subjective fluency. *Acta Psychologica, 98,* 267–290.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Dodson, C. S., Prinzmetal, W., & Shimamura, A. P. (1998). Using Excel to estimate parameters from observed data: An example from source memory data. *Behavior Research Methods, Instruments, and Computers, 30,* 517–526.

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal detection theory and determination of confidence intervals—Rating method data. *Journal of Mathematical Psychology, 6,* 487–496.

Excel 2001 [Computer Software]. (2001). Redmond, WA: Microsoft Corporation.

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

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *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: Human Learning and Memory, 2,* 21–31.

Glanzer, M., Hilford, A., Kim, K., & Adams, J. K. (1999). Further tests of dual-process theory: A reply to Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 522–523.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology; Learning, Memory, and Cognition, 25,* 500–513.

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 61,* 23–29.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355–1369.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace model. *Psychological Review, 95,* 528–551.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory & Language, 33,* 1–18.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 667–680.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

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, and Cognition, 26,* 1534–1555.

Kelley, R. & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27,* 701–722.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

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

Metz, C. E. (1998). Rockit computer program (Computer software). Department of Radiology, University of Chicago. Retrieved from http://www-radiology.uchicago.edu/krl/toppage11.htm

Murdock, B. B. (1965). Signal detection theory and short-term memory. *Journal of Experimental Psychology, 70,* 443–447.

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

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver-operating characteristic curve parameters. *Journal of Mathematical Psychology, 5,* 377–391.

Quamme, J. R., & Yonelinas, A. P. (2001). *A role for familiarity in associative recognition.* Poster presented at the 42nd annual meeting of the Psychonomic Society, Orlando, FL.

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

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

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hikari, 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, and Cognition, 26,* 294–320.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall to reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43,* 67–88.

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

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1499–1517.

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, and Cognition, 24,* 1379–1396.

Swets, J. A. (1986). Indices of discrimination and diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600.

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

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25,* 747–763.

Yonelinas, A. P. (1999a). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1415–1434.

Yonelinas, A. P. (1999b). Recognition memory ROCs and the dual-process signal detection model: Comment on Glanzer, Kim, Hilford, and Adams (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 514–521.

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.

Received June 15, 2001
Revision received March 8, 2002
Accepted March 13, 2002 ■