Item Recognition Memory and the ROC

Andrew Heathcote

The University of Newcastle, Australia

Abstract

Four experiments investigated the effects of study time, study repetition, semantic and orthographic similarity and category length on item recognition memory ROCs (receiver operating characteristics). Analyses of ROC shape rejected Yonelinas's (1994) dual-process model. The normal unequal variance signal detection model provided a better account of the data, except for a small but consistent excess of high confidence errors. Z transformed ROC slope was increased by similarity, category length, and study item repetition, rejecting Ratcliff, McKoon and Tindall's (1994) "constancy-of-slopes" generalization for these variables, but slope was relatively unaffected by massed study time.

Recognition Memory and ZROC

In recent years, ROC (receiver operating characteristic) analysis of recognition memory data has played an increasingly important role in empirical research (e.g., Glanzer, Kim, Hilford, & Adams, 1999a, 1999b; Gronlund & Elam, 1994; Hilford, Glanzer, Kim & DeCarlo, 2002; Kelley & Wixted, 2001; Qin, Raye, Johnson & Mitchell, 2001; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Ratcliff, Van Zandt, McKoon, 1995; Rotello, Macmillan & Van Tassel, 2000; Yonelinas, 1997, 1999a, 1999b; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996; Yonelinas, Kroll, Dobbins & Soltani, 1999) and modelling (e.g., Dennis & Humphreys, 2001; Glanzer, Adams, Iverson & Kim, 1993; McClelland & Chappell, 1998; Murdock, 1998; Shiffrin & Steyvers, 1997; Yonelinas, 1994). ROC analysis is useful because it allows a direct assessment of the distribution of the latent random variable, usually called "familiarity", that is often assumed to subserve recognition decisions.

The ROC is a plot of hit probabilities (H) for old (studied) words against false alarm probabilities (FA) for new (unstudied) words. Points on the ROC are obtained for a set of differing decision criteria, most commonly by asking participants to rate their decision confidence for binary (new versus old) recognition memory judgements. ROC data is usually analysed with a version of signal detection theory (Green & Swets, 1966) that assumes normally distributed familiarity and allows unequal variance for new (σ^2_n) and old (σ^2_o) distributions.

The normal unequal-variance signal-detection (NUS) model predicts that a transformed version of the ROC plot, the ZROC, is linear. The ZROC plots the inverse cumulative normal probability (Z) transformation of hit and false alarm probabilities. Almost all existing quantitative theories of recognition memory also

produce exactly or approximately linear ZROC plots for at least for some of their parameter values. Hence, estimates of the ZROC slope and intercept provide a succinct summary of the data that can, in many cases, be directly mapped to the parameters of a quantitative theory.

Under the NUS model, the slope of a ZROC equals the ratio of new to old familiarity standard deviations, $\underline{S} = \sigma_n/\sigma_o$, and the intercept, $\underline{I} = (\mu_o-\mu_n)/\sigma_o$, where μ_o is the mean familiarity for old items, and μ_n the mean familiarity for new items. Accuracy is usually defined as the standardized distance between the means of the new and old familiarity distributions. The intercept of the ZROC function is, therefore, a measure of accuracy in old standard deviation units. The intercept divided by the slope gives accuracy in new standard deviation units, $\underline{D}' = \underline{I}/\underline{S} = (\mu_o-\mu_n)/\sigma_n$.

For the large class of quantitative theories where familiarity is the sum of *N* random variables (e.g., Humphreys, Pike, Bain & Tehan's, 1989, "global" memory models) or a linear transformation of that sum (e.g., Attention-Likelihood Theory, Glanzer, Adams, & Iverson, 1991) the central limit theorem dictates that the NUS model is exactly correct in the limit of large *N*. The limit is usually holds either by design (e.g., sufficiently long study lists for the global memory models) or by assumption (e.g., a sufficiently large numbers of features and probability of feature marking in Attention Likelihood Theory). In these cases, $\underline{S}$ and $\underline{I}$ can be directly mapped to parameters of the theories. ZROC analysis has become an important tool for testing the detailed predictions of these theories, particularly predictions about ZROC slope (e.g., Glanzer et al., 1993; Ratcliff et al., 1992, 1994, 1995).

More recent quantitative theories of recognition memory do not predict exactly linear ZROC functions, but the empirical pattern of data found with ROC analysis has remained important for testing these theories. Dennis and Humphreys

(2001) fit their model directly to ZROC data. McClelland and Chappell (1998) and Shiffrin and Steyvers (1997) demonstrated that their models could produce approximately linear ZROC functions with slopes less than one, at least for some parameter values. However, the familiarity distributions produced by these theories can deviate substantially from normality, and so the S and I estimates provided by ZROC analysis cannot be directly mapped to model parameters.

Analysis of ZROC shape has been particularly important for another recent model, Yonelinas's (1994) dual-process signal-detection (DPS) theory. DPS postulates that two statistically independent processes underlie recognition memory decisions, a recollection process, which either succeeds or fails with some probability, and a continuous process that is used only if recollection fails. The continuous process follows a normal equal-variance signal-detection model, which produces a linear ZROC with a slope of one. As the probability of recollection increases the ZROC predicted by DPS becomes increasingly nonlinear. DPS contrasts with the theories of McClelland and Chappell (1998) and Shiffrin and Steyvers (1997) in that it cannot produce linear ZROC functions when slope is substantially less than one. Advocates of DPS theory (e.g., Yonelinas, 1997, 1999a, 1999b; Yonelinas et al., 1996; Yonelinas et al., 1999) have sought support from experimental demonstrations of the predicted systematic deviations from ZROC linearity.

Despite intensive research, the interpretation of recognition memory ROCs remains controversial. The following sections review two areas of controversy, about which experimental variables affect ZROC slope, and about the shape of empirical ZROC functions, and, hence, the applicability of the NUS model. In order to test the issues raised, the experiments reported here examine the effects of prototypical

"strength" (study time and study item repetition) and "interference" (similarity and category length) manipulations on ZROC slope and shape.

ZROC Slope

Ratcliff et al. (1992), reporting new experiments and surveying existing data, found that item recognition ZROC slopes were significantly less than one. Assuming the NUS model, these results indicate that the familiarity of studied items is more variable than the familiarity of unstudied items. Although slope decreased with study time at low levels of accuracy, it tended to remain constant with increased study time when accuracy, as indexed by the ZROC intercept, was greater than 0.5. Ratcliff et al. noted that these findings are problematic for global memory models, because they predict either that slope has a constant value of one (e.g., the continuous memory version of TODAM, Murdock & Kahana, 1993), or that it decreases with increasing study time (e.g., SAM, Gillund & Shiffrin, 1984; MINERVA2, Hintzman, 1986). Attention-Likelihood Theory (Glanzer, et al., 1991), also predicts a decrease in slope with study time.

Ratcliff et al. (1994) extended these findings by manipulating accuracy using list length and similarity; once again, they concluded that slope was unaffected for intercepts greater than 0.5, although word frequency was found to affect slope. They summarised their findings with the "constancy-of-slopes" generalization: ZROC slope is unaffected by accuracy above a threshold value, except when accuracy is changed by "materials manipulations", such as word frequency. However, because the constancy-of-slopes generalization requires affirming a null hypothesis, it is difficult to discount objections based on a lack of power.

Recently, Glanzer et al. (1999a) criticised Ratcliff et al.'s (1994) constancy-of-slopes generalization on the basis of new experiments examining word frequency, depth of encoding, study time and study item repetition, along with new analyses of

experiments by Glanzer and Adams (1990) and 33 other experiments. They found that slope generally decreased as accuracy increased, except where accuracy was manipulated by repetition of study items. Confusingly, the largest and closest to significant effect of study on ZROC slope reported by Ratcliff et al. (1992) was produced by repetition of study items.

Although Glanzer et al.'s (1999a) results appear at odds with the constancy-of-slopes generalization, a strong conclusion is difficult to make because accuracy was low in most of the experiments they examined. As noted by Ratcliff et al. (1992), ZROC slope approaches one as the intercept approaches zero both empirically (when study time is short), and according to most recognition memory theories. Given that slope is less than one when accuracy is high, a change in slope is expected when low and high accuracy conditions are compared. Glanzer et al. reported that, in their new experiments, they still found significant decreases in slope as accuracy increased when participants with intercepts less than 0.45 were removed. However, the limit of the constancy-of-slopes generalization to intercepts greater than 0.5 is approximate, and only a minor modification in the threshold would bring Glanzer et al.'s conclusion into doubt. Even for participants retained in the censored analysis, discrimination was still fairly weak, with average intercepts less than one for all but one condition in their third experiment.

Hirshman and Hostetter (2000) reported a significant decrease in slope between conditions with intercepts of 0.72 and 1.24, and stated, "all subjects show discriminability that is substantially above chance" (p. 164). They suggested that their results agree with Glanzer et al.'s (1999a) but differ from Ratcliff et al.'s (1994) because Ratcliff et al. used more study-test lists in their experimental designs. In particular, they suggested that increased inter-list proactive interference in Ratcliff et

al.'s design might have swamped any effects of study on slope. An alternative explanation is that accuracy in Hirshman and Hostetter's experiment was quite low, so that the lower accuracy condition had a higher slope because it contained low accuracy subjects with small intercepts.

Glanzer et al. (1999a) also criticised Ratcliff et al.'s (1994) limitation of the constancy-of-slopes generalization to other than materials manipulations on the grounds that it is unprincipled: "no reason is given why 'type of materials' and 'strength manipulations' both of which affect accuracy, should have differed in their effect on slopes." (p. 509). Ratcliff et al. did not find a significant effect on slope of similarity between new and old words (although close, $\underline{p} = 0.059$), but similarity might be argued to be a materials manipulation. Once again, power is a problem, not only in terms of the marginally significant result, but also because semantic similarity was manipulated. Semantic similarity often has a much weaker effect than orthographic similarity on recognition memory (e.g., Gillund & Shiffrin, 1984). Further, Ratcliff et al. used a category length (number of similar items in the study list) of only one. Item noise theories (see Dennis & Humphreys, 2001), which include the global memory models, as well as the McClelland and Chappell (1998) and Shiffrin and Steyvers (1997) theories, predict stronger similarity effects with increasing category length. Hence, Ratcliff et al.'s demonstration of a null effect on slope would have been more convincing if they had used longer categories.

Shiffrin, Huber and Marinelli (1995) investigated the effects of both semantic and orthographic similarity, and category length and strength, on recognition memory. Results were generally in accord with the differentiation version of the SAM theory (Shiffrin, Ratcliff & Clark, 1990), which predicts that similarity and category-length affect ZROC slope, but that category strength does not. However, the evidence for an

increase in the hit rate with category length (predicted by SAM) was weak and inconsistent, despite a clear increase in the false-alarm rate. A ceiling effect on hit rates was unlikely as accuracy was low, so Shiffrin et al. suggested a dual-process explanation (e.g., Atkinson & Juola, 1974; Mandler, 1980). They speculated that hits could sometimes result from recollection but that recollection becomes less effective as category length increases. As familiarity-based increases in hit rate with category length trade-off with decreased hit rates due to recollection, the effect of a category length on overall hit rate is weakened.

Several features of Shiffrin et al.'s (1995) design were not conducive to finding strong effects of category length. Categories were embedded in very long lists in order to avoid participants becoming aware of the category length manipulation and adjusting their decision criteria. For the SAM model, category length as a proportion of list length determines the magnitude of the category-length effect. Hence, it is not surprising that Shiffrin et al. found only weak effects of category length on accuracy, although the effects were significant. Another problem is that Shiffrin et al.'s design did not yield a sufficient number of observations for analyses of individual-participant ZROCs. Fits of the SAM model to averaged ZROCs were swamped by variance due to between-subjects differences and variance due to long lists. As a result, the fits were unable to capture even their weak effects of category length on accuracy, because extremely high estimates of old familiarity variance were necessary to accommodate the average ZROC slope.

In summary, the effect on ZROC slope of manipulations that affect accuracy are controversial, and experimental evidence is inconsistent. Ratcliff and colleagues have found that slope is unaffected by study time, although repetition produced a marginal effect. Glanzer and colleagues found clear effects of massed study time, but

not spaced study repetitions. Where there is agreement that slope and accuracy change together (e.g., materials manipulations such as word frequency) no explanation is evident as to why these cases differ. Ratcliff et al. (1994) found that similarity, which appears to be a materials manipulation, produced at best a marginal effect on slope, contradicting the predictions of SAM. In contrast, Shiffrin et al. (1995) supported SAM's general account of similarity and category length effects. However, they could not obtain satisfactory fits of the SAM model, and needed to invoke a qualitatively specified dual-process mechanism. The next section examines a quantitative dual-process theory, which Yonelinas (1994) proposed as an explanation of results about both ZROC slope and ZROC shape.

ZROC Shape

Yonelinas (1994) proposed dual-process signal detection (DPS) theory as a unifying account of experimental results on ZROC slope. DPS theory predicts linear ZROCs with unit slope when recollection is absent. As the probability of recollection increases, ZROC functions become progressively more concave upwards (i.e., U shaped), and the slope of a straight-line fit to the ZROC function becomes progressively less than one. Consequently, the shape of ZROC data can be used to adjudicate between DPS theory and the NUS model, particularly when slope estimates are less than one. Because DPS theory and the NUS model both have two free parameters, and hence similar flexibility, they can also be compared by goodness-of-fit to ZROC data.

Yonelinas (1994) used DPS theory to provide a principled explanation of Ratcliff et al.'s (1994) constancy-of-slopes generalization. A simulation demonstrated that simultaneous increases in recollection and familiarity result in an increase in ZROC intercept with no change in slope, whereas a change in recollection alone increased the intercept and decreased the slope. Yonelinas hypothesised that

manipulations affecting interference, in which he included lag and delay manipulations (Donaldson & Murdock, 1968; Gehring, Toglia & Kimble, 1976) as well as materials manipulations, change only the probability of recollection, whereas strength manipulations, such as study time, cause simultaneous changes in both the signal detection and recollection components. Hence, DPS provides a potential explanation of the constancy-of-slopes generalization.

The only evidence for concave upward ZROCs in item recognition comes from Yonelinas et al. (1996) in conditions that may have promoted recollection, such as requiring source recall. Glanzer et al. (1999a) found virtually no evidence for concave upward functions in their extensive review of ZROC results for item recognition. They concluded that: "[upward] concavity is not a general characteristic…[and] does not present a general explanation of [Z]ROC slopes" (p. 512). Ratcliff et al. (1995) also found no support for concave upward deviation in the data of Ratcliff et al. (1994), and that estimates of recollection probability obtained by fitting Yonelinas's (1994) model changed in an unreasonable way for this data. Yonelinas (1994) found similar unreasonable changes in recollection estimates but attributed them to floor and ceiling effects. Ratcliff et al. (1995) note that floor and ceiling effects cannot provide a similar explanation for their findings.

Yonelinas (1999a) countered Glanzer et al.'s (1999a) criticism of DPS theory by showing that it provided good fits (his Figure 3) to ten of Glanzer et al.'s averaged data sets. However, this defence is open to criticism because a good fit does not, by itself, provide strong evidence for a model (Roberts & Pashler, 2000). This is particularly true of ROC data, as both observed ROC curves and ROC curves predicted by any reasonable theory are constrained to always be non-decreasing. The DPS theory and the NUS model are also exactly equivalent when ZROC slope equals

one, with differences only emerging as slope decreases. Hence, the fit of these models would be expected to be similar.

To check Yonelinas's (1999a) findings, the DataThief[1] software was used to measure the 10 data sets from his Figure 3 and fits of both DPS theory and the NUS model obtained as described by Yonelinas (i.e., least squares on probabilities). Figure 1 displays the data from Glanzer et al.'s (1999a) Experiment 1 and the DPS fits. As reported by Yonelinas, the average R^2 for the NUS model was superior to that of DPS only in the fourth decimal place. However, the NUS model fit better than DPS for every one of the 10 curves, a significant result by a binomial test. As shown in Figure 1, where ZROC slope was shallow, a concave upward bend was evident in the DPS fits, but where slope was near one, DPS fits were close to linear[2] (c.f., Ratcliff et al., 1995). The correlation of the difference in R^2 between the two models and ZROC slope was also significant ($r = 0.67$, $p = 0.03$), indicating that the fit of DPS theory became worse as ZROC slope decreased.

Ratcliff et al. (1994) proposed an alternative explanation of nonlinear ZROCs, where responses based on the NUS model are contaminated with guesses. They found that even a small proportion of guesses that were uniformly distributed across confidence levels produced a concave downward (i.e., inverted U) ZROC. Because guessing responses are assumed to be relatively rare, they cause only small changes in response probabilities. However, their effect is still appreciable for rare responses (e.g., high confidence false alarms and misses in accurate responding), where the proportional change due to guessing is large. The effect of guessing on rare responses is also emphasised in ZROC plots by the highly nonlinear nature of the Z transformation for probabilities near zero or one.

DPS theory predicts an excess of high confidence hits, resulting in the strongest deviations from linearity occurring upward and on the left of the ZROC function. Ratcliff et al.'s (1994) guessing model, in contrast, produces downward deviations on the right side of the ZROC and rightward deviations on the left side of the ZROC. It is difficult to compare the two models using goodness-of-fit because the guessing model has one more parameter than DPS theory. However, the distinct patterns of deviations from linearity mean that ZROC shape can be used to compare them, by counting the number of cases that display concave upward versus downward shape (cf. Glanzer et al., 1999a).

In summary, Yonelinas's (1994) DPS theory may, in principle, provide an explanation of exceptions to the constancy-of-slopes generalization. However, direct evidence for DPS theory in item recognition, concave upward ZROCs, is lacking. As ZROC slopes less than one are commonly seen in item recognition paradigms, and DPS theory must assume concave upward ZROC functions to account for slopes less than one, its explanation of the constancy-of-slopes generalization is open to question.

Overview of Experiments

The present experiments test both Yonelinas's (1994) DPS theory and Ratcliff et al.'s (1992) constancy-of-slopes generalization in item recognition. All experiments employed a prototypical method of manipulating interference: item similarity. Experiments 1 and 2 investigated semantic and orthographic similarity respectively, and Experiment 3 investigated the category-length effect for orthographic similarity. Similarity and category length were manipulated as within-subjects variables. DPS theory predicts that similarity and category length will affect ZROC slope, because they are interference manipulations, whereas the constancy-of-slopes generalization predicts that slope will not be affected by similarity and category length.

Experiments 1 and 2 also manipulated study time. Manipulation of accuracy through study time is the prototypical case where Ratcliff et al.' (1992) constancy-of-slopes generalization should hold. DPS theory also predicts that ZROC slope will not change if accuracy is manipulated by study time. Study time was manipulated in two ways between subjects: by repeating study of the same item at spaced intervals (the spaced condition) and by increasing the time of a single study presentation (the massed condition). Comparison of these two manipulations of study time is of interest because previous evidence suggests that they may have different effects on ZROC slope.

Memory is better for spaced than massed study when the study-test lag is moderate or large compared to the lag between spaced repetitions (Glenberg, 1976), as was he case in the present experiments. In recognition memory, the spaced advantage is usually attributed to deficient study of the massed repetitions. Greene (1989) found that the spacing effect disappeared under incidental learning conditions. He suggested that the spacing effect in cued-memory tests such as item recognition is due to voluntary rehearsal; participants assume that a massed repetition's greater familiarity indicates that it requires less rehearsal. However, Challis (1993) found that the spacing effect could occur with incidental learning, as long as the orienting task required lexical/semantic processing. He suggested that the spacing effect in recognition is due to an involuntary reduction in the study of massed repetitions due to lexical/semantic priming from earlier presentation. The intentional study conditions used in the present experiments make it likely that participants will engage in lexical/semantic processing, and hence produce a spaced advantage.

Several aims constrained the design of the experiments. Most importantly, ZROCs had to be estimable for each individual participant and condition. Averaged

ZROCs confound between-subject variance and intrinsic variance in familiarity. It is the intrinsic variance of familiarity that is the subject of recognition memory theories. Most of the analyses of Glanzer et al.'s (1999a), Ratcliff et al. (1992), and Ratcliff et al. (1994) used averages over participants. For these analyses, changes in slope may have been produced by changes in individual differences rather than changes in familiarity variance. Conversely, a failure to detect a change in slope may have been caused by an interaction of individual differences and familiarity variance effects.

A second design constraint was that good accuracy was necessary in all conditions, so that finding an effect on ZROC slope could not be attributed to the inclusion of a low accuracy condition. This aim was facilitated by the use of short study lists. Short lists are also desirable because, as demonstrated by Gronlund and Elam (1994), long lists yield ZROC slopes close to one. Within the global memory model framework, this occurs because differences between new and old familiarity variance are swamped in long lists by variance due to small matches from the large number of study list items. The shallow slopes that occur with short lists provide a powerful test of Yonelinas's (1994) DPS theory, because it cannot predict such slopes without easily detectable nonlinearity in ZROCs (Ratcliff, et al., 1995).

Experimental effects and sample sizes had to be large so that a failure to find ZROC slope differences could not be attributed to a lack of power. In pilot testing, large effects were achieved by doubling study time and using one versus two repetitions. A large effect of similarity was found when one third of the study list consisted of similar words, the level adopted in Experiments 1 and 2. Large effects of category length were achieved in Experiment 3 by using between one-sixth and one-half similar words in study lists.

Strong effects of similarity also required careful selection of words. Similar words were drawn from pools of 24 words with high mutual similarity (see Appendix). Semantically similar lists consisted of the 24 most frequently produced words in response to category labels for a sample of 620 Australian participants (Casey, 1988). Inspection of these lists reveals that they are largely similar to standard lists, but also contain some unique local usages. These usages were appropriate for the participants in the present experiments as Casey's sample was drawn from the same population (university students from New South Wales in Australia). Orthographically similar lists, with an average of more than 50% identical letters in corresponding positions, were produced by a computer-aided search of the MRC2 database (Coltheart, 1981).

Analysis Techniques

Estimates of the NUS model for each individual participant were obtained using a maximum likelihood procedure that estimates parameters for any number of within-subject conditions simultaneously (Kijewski, Swensson, & Judy, 1989), rather than being limited to pairs of conditions (c.f., Dorfman & Alf, 1969; Grey & Morgan, 1972; Ogilvie & Creelman, 1968). Simultaneous estimation provides parameter values relative to a reference condition, which by assumption has a familiarity distribution with a mean of zero and standard deviation of one. The simultaneous estimation method also assumes that the same decision criteria are used for all conditions. Given these assumptions are true, the simultaneous estimation procedure is more efficient than procedures based on pairwise estimation (Kijewski et al., 1989).

The large numbers of similar words in study lists may have made the similarity manipulation obvious to some participants. Hence, they may have used different decision criteria for similar and dissimilar words, and for lists having different category lengths. To avoid violating assumptions, simultaneous maximum

likelihood estimation was carried out separately for similar and dissimilar words and for different category lengths. Within these conditions, parameters for new, weak and strong conditions, which must have the same decision criteria, were estimated simultaneously, using the new condition as a reference.

The primary method of presenting the goodness-of-fit results for the NUS model is graphical: plots of Z scores for each condition on the ordinate against the Z scores for the reference distribution. However, simply plotting average observed Z scores and Z scores obtained by averaging parameters from individual model fits can be quite misleading when missing values are present. Some participants had missing values because the Z transformation is undefined for probabilities of zero or one, a common occurrence for the incorrect high confidence rating categories when, as in the present experiments, accuracy was high[3].

The solution adopted here was to plot the data as average deviations from the Z values predicted by the average model (i.e., the model with parameters that were the average of each participant's parameter estimates). This provides an accurate reflection of average deviations from the NUS model. Plots include 95% normal confidence intervals (i.e., $\pm 1.96 SD(d_i)/\sqrt{n}$, where n is the number of participants and d_i is the i^{th} participant's deviation) for both Z(H) (vertical bars) and Z(FA) (horizontal bars). When inspecting plots, comparison of the NUS model predictions (solid points) and the confidence intervals are useful for determining the reliability of deviations in Z(H), Z(FA), or both. It is important to note that the NUS model may accurately fit Z(H) but fail for Z(FA), or vice versa, for any given point. For example, most of the large deviations observed in the following experiments were for (rare) high confidence errors, whereas very little deviation was seen for the corresponding (common) high confidence correct responses.

Goodness-of-fit was also quantified using a χ^2 test. Simulation studies have shown that this test rejects the null hypothesis too often (Metz, 1988, cited in Kijewski, et al., 1989) unless the expected number of observations is greater than 5 in all rating categories. This was often not the case for high confidence errors, so a tight rejection criterion of $\underline{p} < 0.01$ was selected. The results for the χ^2 tests were also used to select cases for further analysis of the pattern of deviation, using the quadratic regression method of Glanzer et al. (1999a). Any ZROC with significant deviation according to a liberal criterion of $\underline{p} < .1$ was subject to quadratic regression analyses. The signs of quadratic regression coefficients were used to assess whether individual ZROCs were concave upward, as predicted by DPS theory, or concave downward, as predicted by Ratcliff et al.'s (1994) guessing model. Unless otherwise specified, significance of inferential tests was assessed at the 0.05 level.

Accuracy was analysed using both the intercept (I) and intercept divided by slope (D') estimates. Analyses of I have been reported in most previous ROC studies and so are provided for comparison. Analyses of D' are reported as slope, and hence the relative values of I and D', may vary across conditions. Measures combining estimates of old and new variance, such as $d_a = (\mu_o - \mu_n) \Big/ \sqrt{(\sigma_n^2 + \sigma_o^2)/2}$ and $d_e = (\mu_o - \mu_n)/((\sigma_n + \sigma_o)/2)$, fall between I and D' estimates, regardless of slope, so analysis of both I and D' provide a broad characterization of effects on accuracy.

Experiment 1

Experiment 1 examined the effect of semantic similarity and study time, weak (3 seconds/pair) versus strong (6 seconds/pair), as within-subject variables. Study words were presented in pairs and subjects instructed to create associations between pair members in order to focus study on the currently presented pair. Spaced versus massed presentation of strong items was a between-subject variable. Participants

simultaneously indicated whether a test word was new or old and their confidence on a six-point scale.

Methods

Participants

The 64 participants were students at the University of Newcastle, Australia, who volunteered to participate in a one-hour session. Half of the participants were randomly allocated to the massed study condition and half to the spaced study condition.

Apparatus and Stimuli

An IBM-PC clone with a colour monitor presented stimuli and recorded responses. Confidence ratings were obtained using the "z", "x", "c", ",", ".", and "/" keys, indicating "Sure New", "Probably New", "Possibly New", "Possibly Old", "Probably Old", and "Sure Old" respectively. During testing, the words "Sure", "Probably", "Possibly", "Possibly", "Probably" and "Sure" were arrayed across the bottom of the screen with the words “New” and “Old” above “Probably” on the left and right respectively. Timing and the synchronization of stimulus presentation with screen refresh were achieved using Heathcote's (1988) programs.

Similar stimuli were drawn from sets of 24 words from 18 categories (see Appendix). The set of dissimilar stimuli consisted of 776 words selected so that they did not come from any of the categories used for similar words. Forty of the dissimilar words were used for practice and the remaining 736 for testing. The 736 dissimilar words were approximately matched to the similar words on their average natural log Kucera-Francis and Thorndike-Lorge counts (taken from the MRC2 database, Coltheart, 1981). Non-zero Kucera-Francis counts were available for all dissimilar words and all but 36 similar words. Non-zero Thorndike-Lorge counts were available for all but 16 dissimilar words and 31 similar words. Only 7 similar words had zero

Kucera-Francis and Thorndike-Lorge counts. The average natural log frequency was 3.89 for dissimilar words and 3.54 for similar words.

Procedure

Participants studied and were tested on one practice list and 18 experimental lists of 20 words. Each study list took 36 seconds to present. Study words were presented as pairs in white text in the middle of a black screen. Participants were instructed to form associations between pairs of words and to rehearse only the pair of words that was on the screen during study-list presentation. Pair members were always both from the similar word set or both from the dissimilar word set.

The first and last pair in each study list was presented for three seconds. One of these four "buffer" words was randomly selected for testing but the response to it was not recorded. On average, half of the pairs presented in the middle 30 seconds of the study list were similar and half were dissimilar. For half of the massed-study lists, four "strong" pairs were presented for 6 seconds each, followed by two "weak" pairs presented for 3 seconds each. For the other half of the massed lists, four weak pairs were presented for 3 seconds each followed by three strong pairs presented for six seconds each. For spaced study, each pair was presented for three seconds and no two identical pairs followed each other. Each study list had two, three or four "strong" pairs, which were presented twice, and six, four, or two weak pairs respectively, which were presented once. The first presentation of the strong pairs always occurred in the first four pairs after the first buffer pair. Apart from these constraints the presentation order was random.

Following study, the screen was blank for one second and then a new screen appeared asking participants to initiate testing by pressing a space bar. Test words were presented in random order and consisted of the 19 new words not seen during study, a buffer word, and 13 words selected at random from the set of non-buffer

words presented during study. For each participant and study-test cycle, old and new words were randomly selected without replacement from a list of similar words and from the pool of dissimilar words.

A test screen displayed the test word at the centre of the screen, and a visual-analogue representation of the confidence rating scale at the bottom of the screen. When a participant responded, the corresponding word on the rating scale flashed for half a second. New responses were made with the left hand and old responses with the right hand. The next test screen was presented one second after the previous response. At the end of a study-test cycle, the screen went blank for two seconds. A new screen then appeared asking the participant to press any key to begin the next study-test list. Participants were encouraged to take a break at this point if they desired.

At the start of the session, all participants studied, and were tested on, a practice list constructed from the same set of 40 dissimilar words. The response procedure was described to participants and they were encouraged to make use of all confidence ratings and to respond as rapidly as was compatible with maintaining high accuracy. Participants were not informed about the manipulation of word similarity. The procedure for the practice list was identical to that for the experimental lists, except that, following the practice test, participants were given feedback about the number of their responses in each confidence category. The experimenter discussed this feedback with the participant and emphasized the need to use all confidence categories. Participants were instructed to achieve this aim by adjusting their definitions for each confidence category (e.g., by reserving “sure” responses for occasions on which they were “absolutely and completely sure” if they tended to neglect low confidence responses). They were discouraged from randomly responding or choosing rarely used confidence categories without reference to their actual feeling

of confidence. At the conclusion of the session participants were thanked for their participation and any questions about the experiment answered.

Results and Discussion

Figure 2 illustrates the fit of the unequal-variance normal-distribution (NUS) model to the confidence data. Only 2 of 64 fits in the spaced condition and 4 of 64 fits in the massed condition were rejected by the χ^2 test at the 0.01 level. The most pronounced deviations occurred for the most infrequent types of responses: high confidence errors (i.e., hits on the upper right of the figure and false alarms on the lower left of the figure). Participants produced more high confidence errors, to both new and old words, than predicted by the NUS model. As previously discussed, a small excess of high confidence errors is consistent with an underlying NUS process contaminated by a small proportion of guesses.

Two factors mitigate the deviations from the NUS model's prediction of ZROC linearity. First, missing values were common for the deviant ratings because many participants did not make high confidence errors. Second, the deviations are exaggerated in the plots because the Z transform amplifies small probability differences near zero or one. At most, these deviations represent a difference between expected and observed probabilities of 0.009 and a corresponding frequency difference of only 0.54 responses (for strong similar hits in the massed condition, Figure 2b). In short, the deviations for high confidence errors are at least partially due to floor and ceiling effects that would be expected given the good overall discrimination between old and new words.

One aspect of the observed misfit not attributable to guessing and/or floor and ceiling effects was evident at the boundary between new and old responses in the strong massed similar condition. The difference between expected and observed

probabilities was 0.023 with a corresponding frequency difference of 1.34 responses. This pattern of misfit was not evident in any other condition of any experiment reported here, so its origin is unclear.

Where deviation occurred at the individual level, it followed a consistent concave downwards pattern. For cases with significant misfit according to the χ^2 test at the 0.1 level (the liberal criterion used to select individual ZROCs for polynomial regression analysis), negative quadratic coefficients indicative of concave downwards deviation were found for 24 of the 28 polynomial regressions of Z(H) against Z(FA). Hence, the observed minor deviations from a linear ZROC function are consistent with a small percentage of guessing responses for some participants.

In summary, there was no support for the pattern of deviations predicted by Yonelinas's (1994) DPS theory, despite the fact that the slopes were much less than one, so deviations should have been large enough to detect. In particular, the critical high confidence hit results were extremely well predicted by the NUS model and virtually no concave upward deviation was found in individual ZROCs.

ZROC Parameters

Figure 2 displays the intercept, slope, and D' (intercept divided by slope) estimates calculated by averaging the estimates from individual fits. Similar slope was greater than dissimilar slope, $F(1, 62) = 9.68$, MSE = 0.084. The interaction between similarity and study distribution just achieved significance, $F(1, 62) = 4.02$, $p = 0.049$, as the difference between dissimilar and similar slope was greater for weak than strong items. These results cast doubt on Ratcliff et al.'s (1994) constancy-of-slopes generalization for the effect of similarity.

The effect of study time on slope was weaker than the effect of similarity on slope. On average, strong slope was less than weak slope for massed study, but the order was inconsistent for similar and dissimilar words (Figure 2b vs. Fig. 2a). For

spaced study, in contrast, strong slope was greater than weak slope for both similar and dissimilar words (Figure 2d vs. Fig. 2c). The different effects of massed and spaced study on slope was reflected in a significant interaction between study distribution and study time, $F(1, 62) = 6.21$, MSE = 0.024.

Separate analyses of the massed and spaced conditions found a significant main effect of study time on slope for the spaced condition, $F(1, 31) = 4.86$, MSE = 0.05, but not for the massed condition, $F(1, 31) = 1.43$, MSE = 0.024. No other effects were significant, except the interaction between study time and similarity in the massed condition, $F(1, 31) = 4.66$, MSE = 0.021, reflecting the greater effect of study time on slope for similar than dissimilar words. A t-test of similar word data from the massed condition (Figure 2b) found that weak slope was significantly greater than strong slope ($t(31) = 2.33$, SE = 0.038).

The results for spaced study reject Ratcliff et al.'s (1994) constancy-of-slopes generalization because spaced study caused a significant increase in slope. The global memory models cannot account for these results, as they predict that slope should either decrease (SAM, MINERVA2) or stay constant (TODAM). The results for massed study are more ambiguous. For similar words, massed study decreased slope as predicted by SAM and MINERVA2, but for dissimilar words it had little effect. In all cases slope was clearly less than one, contradicting the predictions of TODAM.

For ZROC intercepts, the interaction between study distribution and study time was highly significant, $F(1, 62) = 34.2$, MSE = 0.136, with spaced strong study greater than massed strong study. The same was true for the D' measure, $F(1, 66) = 15.0$, MSE = 0.212. The effect of similarity on the intercept was surprising; the similar intercept was greater than the dissimilar intercept, $F(1, 62) = 19.2$, MSE = 0.193. For the D' measure, in contrast, similar was less than dissimilar, except for

strong massed study. Although the main effect of similarity on D' was not significant, F(1, 62) =2.09, MSE = 0.936, the three-way interaction between all factors was just significant, F(1, 62) = 4.09, MSE = 0.212, p = 0.049, consistent with dissimilar accuracy being greater than similar accuracy in all but the strong massed condition.

In summary, accuracy increased with study time, with a greater increase for strong spaced than strong massed study. Similarity had opposite effects on the two accuracy measures derived from the ZROC fits, reflecting the strong effect of similarity on ZROC slope. ZROC intercepts generally indicated better accuracy for similar than dissimilar items, whereas D' generally indicated the opposite ordering. Overall, accuracy was high, with the median intercept being 2.01. Hence, the observed changes in ZROC slope were not due to the inclusion of low accuracy data.

Experiment 2

Experiment 2 was almost identical to Experiment 1, except that similarity was manipulated orthographically.

Methods

Participants

The 68 participants were students at the University of Newcastle, Australia, who volunteered to participate in a one-hour session. Half of the participants were randomly allocated to the massed conditions and half to the spaced condition.

Apparatus and Stimuli and Procedure

The same apparatus was used as in Experiment 1 but with different stimuli. The stimuli consisted of 36 sets of 56 words each, 24 words having high mutual orthographic similarity and 32 words having low mutual similarity, and low similarity to the 24 highly similar words. Words in each set had the same number of letters, with 9 five-letter, 12 six-letter, 9 seven-letter and 6 eight-letter sets (see Appendix for a full listing). Orthographic similarity was measured by "overlap": the percentage of

identical letters in the same positions. The similarity of each word to all other words in its list was calculated by the overlap measure and averaged for each set of words. Pairs of similar words had an average 58.5% overlap, pairs of dissimilar words 2.9% overlap, and pairs of similar and dissimilar words 0.7% overlap. All words had at least one occurrence in the Kucera-Francis counts (taken from the MRC2 database, Coltheart, 1981) and similar and dissimilar words were equated on average natural log word frequency (2.076 for similar words and 2.099 for dissimilar words).

The experimental procedure was identical to Experiment 1, except for one detail of study-test list construction. The 18 experimental lists for each participant were drawn randomly without replacement from the pool of 36 word sets. Each study-test list was constructed by random selection from the 24 similar and 32 dissimilar words in each set, rather than dissimilar words being drawn from a common pool as in Experiment 1.

Results and Discussion

Figure 3 illustrates the fit of the NUS model. Significant misfit (at the 0.01 level) occurred for only 2 of the 68 fits in the massed condition, in both cases for dissimilar words, and for 4 of the 68 fits in the spaced condition, twice for dissimilar and twice for similar words. Misfit was mainly evident in the spaced condition, where dissimilar words had an excess of high confidence false alarms (Figure 3c), whereas similar words had an excess of high confidence misses (Figure 3d). The misfit was small, being at most 0.006 in probability and 0.4 in response frequency. As in Experiment 1, the pattern of deviation was systematically concave downwards. For cases with significant χ^2 misfit at the 0.1 level, 22 of the 26 had negative quadratic coefficients consistent with Ratcliff et al.'s (1994) guessing model rather than DPS theory.

As in Experiment 1, no evidence was found to support the DPS theory explanation of ZROC results, despite the fact that slopes were, in most cases, less than one. The NUS model accurately predicted high confidence hits and virtually no concave upward deviation was found in individual ZROCs. Some concave downward deviation was observed in a few cases, consistent with a small percentage of guessing responses for a minority of subjects.

ZROC Parameters

Figure 3 displays D', intercept and slope measures calculated by averaging the estimates from individual fits. Similar slope was greater than dissimilar slope, $F(1, 62) = 69.6$, MSE = 0.048, with the difference being greater (around 0.2 on average) than in Experiment 1 (around 0.1 on average). These results confirm and extend the conclusion that Ratcliff et al.'s (1994) constancy-of-slopes generalization does not hold for the effects of similarity.

The effect of study time on ZROC slope followed the same pattern as in Experiment 1. On average, strong slope was less than weak slope for massed study, but the order was inconsistent for similar (Figure 3b) and dissimilar (Figure 3a) words, whereas strong slope was greater than weak slope for spaced study of both similar (Figure 3d) and dissimilar (Figure 3c) words. The interaction of study distribution and study time was significant, $F(1, 62) = 6.06$, MSE = 0.020. Separate analyses of massed and spaced conditions found no main effect of study on massed slope, $F < 1$, but a significant effect for spaced study, $F(1, 33) = 6.72$, MSE = 0.022. No other effects on slope were significant.

As in Experiment 1, the results for spaced study reject Ratcliff et al.'s (1994) constancy-of-slopes, because spaced study had a significant effect on slope. They also reject the global memory model predictions, because spaced study increased rather than decrease slope. The results for massed study support the constancy-of-slopes

generalization. For similar words, there was a trend for slope to decrease with study, but the difference was not significant according to a t-test, $t(33) = 1.81$, SE = 0.028. For dissimilar words, there was a trend for slope to increase with study, but again the difference was not significant according to a t-test, $t(33) = 1.13$, SE = 0.026.

For ZROC intercepts, the interaction between study distribution and study time was highly significant, $F(1, 66) = 39.6$, MSE = 0.068, with spaced strong study being more accurate than massed strong study, as in Experiment 1. The same was true for the *D'* measure, $F(1, 66) = 10.4$, MSE = 0.092. In contrast to Experiment 1, the effect of similarity on both accuracy measures, intercept and *D'*, was consistent; dissimilar was greater than similar for the intercept, $F(1, 66) = 49.6$, MSE = 0.112 and, for *D'*, $F(1, 66) = 183.9$, MSE = 0.304.

Overall, the effect of study on accuracy followed the same pattern as in Experiment 1, with a greater increase for spaced than massed study. However, the effect of orthographic similarity differed from the effect of semantic similarity: semantic similarity did not have a consistent effect on accuracy, whereas accuracy for orthographically similar words was clearly less than accuracy for dissimilar words on both I and D' measures. Overall accuracy was not as high as in Experiment 1, the median intercept for Experiment 2 being 1.35, but it was still sufficient so the observed changes in ZROC slope were unlikely to be due to the inclusion of low accuracy data.

The increase in slope with spaced study is problematic for the constancy-of-slopes generalization. Combined analysis of the spaced study slopes from Experiments 1 and 2 revealed a highly significant increase of 0.08 for strong compared to weak study, $F(1, 64) = 10.9$, Adjusted MSE = 0.035. A combined analysis of massed study slopes from Experiments 1 and 2, in contrast, did not show a

significant main effect of study, $\underline{F}(1, 64) = 1.79$, Adjusted MSE = 0.017, but did show a highly significant interaction between similarity and study time, $\underline{F}(1, 64) = 8.59$, Adjusted MSE = 0.018. The interaction reflects a small increase in slope of 0.03 for dissimilar words, and a larger decrease in slope of 0.07 for similar words. The small effect for dissimilar words is consistent with the constancy-of-slopes generalization, whereas the larger decease for similar words is not.

Experiment 3

As previously discussed, Shiffrin et al. (1995) investigated the effects of category length using semantic and orthographic similarity, but, due to design constraints, ZROC analysis was only carried out on group data. Experiment 3 examines the category length effect in a design with sufficient observations per participant and condition to allow analysis of each participant's ZROC results. Category length was manipulated using orthographic similarity, as it was previously shown to produce larger effects than semantic similarity. Large effects were also likely because the category lengths were manipulated over a wide range, from 1/6 to 1/2 of list length.

The original (Gillund & Shiffrin, 1984) and differentiation versions (Shiffrin, Ratcliff & Clark, 1990) of the SAM theory predict similarity and category length effects. Although space does not allow presentation of the details (a copy of the proofs may be obtained from the author), it can be shown analytically that SAM predicts: i) decreased accuracy (as measured by both $\underline{I}$ and $\underline{D}$), and increased slopes for similar compared to dissimilar items, and ii) that dissimilar items are unaffected by category length, whereas for similar items accuracy decreases, and slope increases towards one, as category length increases. The mechanisms underlying the category length effect are related to the mechanisms causing list length effects in the global memory models,

as described by Gronlund and Elam (1994). Familiarity variance for new and old items increases with both category and list length such that the ratio of new to old variance, and hence slope, increases towards one.

Methods

Participants

The 34 participants were volunteers from the University of Newcastle, Australia, who participated in two one-hour sessions on different and usually successive days.

Apparatus and Stimuli and Procedure

The same apparatus and stimuli were used as in Experiment 2. In addition to the 36 mixed similar and dissimilar word sets used in Experiment 2, 12 pure dissimilar word sets were also used, 3 with five-letter words, 4 with six-letter words, 3 with seven-letter words and 2 with eight-letter words. The pure dissimilar sets consisted of 40 words. Pairs of words from the pure dissimilar sets had an average 4.7% overlap.

Study words were presented in pairs, study lists took 36 seconds to present, and 33 words were tested, as in Experiment 2. However, the number of buffer pairs was doubled, with two buffer pairs at the beginning of the study list and two at the end. Buffer pairs were always made up of dissimilar words. One of the eight buffer words was randomly selected and tested, but the response to it was not recorded. The central 24 seconds of the study list consisted of the presentation of eight different pairs of words for three seconds each. All of these sixteen old words were tested along with one buffer word and sixteen new words. Either none, two, four or six of the eight central old pairs were made up of similar words, with the remainder being dissimilar word pairs. Pairs were presented in random order. Hence, category length, the number of similar words in a study list, could be zero, 4, 8, or 12.

Twelve study lists were presented at each category length. Each non-zero category-length list was created by randomly sampling from the similar and dissimilar words in one of the 36 mixed word sets. Word sets were randomly allocated to category length conditions for each participant. The zero category length lists were created by randomly sampling from the pure dissimilar lists. Lists of different types were presented in random order. Of the 24 lists presented in each experimental session, 18 were mixed lists, 6 from each category length, and 6 were pure dissimilar lists. The allocation of 24 word sets to each of the two sessions was randomised for each participant with the constraint that approximately equal numbers of sets of each letter length were used in each session. The experimental procedure was the same as for Experiments 1 and 2, except that two sessions lasting approximately one hour each on different days were employed. Practice was performed only on the first day.

Results and Discussion

Figure 4 illustrates the fit of the NUS model. Overall, it accurately characterized most of the data. Only 9 of 238 cases had significant misfit, at the 0.01 level, according to χ^2 tests. Misfit was evenly distributed across conditions, with the most misfits in one condition occurring for 3 of the 34 fits for dissimilar words and category length 8.

The most visually striking deviation in Figure 4 is for high confidence false alarms to similar words in category length 12 lists (Figure 4d). However, the corresponding probability difference was less than 0.001, or less than 0.1 responses. As in previous experiments, the pattern of deviation, while small, was systematically concave downward. For the 19 cases with significant misfit at the 0.1 level, 12 had negative quadratic coefficients indicative of concave downward deviation. As in earlier experiments, the shape of the deviations, and the excellent fit of the NUS

model to high confidence hits, supports Ratcliff et al.'s (1994) guessing model, rather than the recollection effects predicted by DPS theory, as responsible for the slight misfit.

ZROC Parameters

Figure 4 displays the D', intercept and slope measures calculated by averaging the estimates from individual fits. Category length did not effect dissimilar word slope, $F(3, 99) = 2.04$, MSE = 0.087, but had a marginally significant effect on slope for similar words, $F(2, 66) = 2.52$, MSE = 0.037, $p = 0.088$. The trend for similar words was due to an increase in slope for the longest category length. A post-hoc t-test comparing the slope for category length 12 (Figure 4d) with the average slope for the other two category lengths (Figures 4b and 4c) was highly significant, $t(34) = 2.81$, SE = 0.032, even when allowing for the post hoc nature of this test.

The intercept for dissimilar words did not change significantly with category length, $F(3, 99) = 2.04$, MSE = 0.087, but the effect of category length on D' was significant, $F(3, 99) = 2.89$, MSE = 0.474. For similar words, category length had a significant effect on the intercept, $F(2, 66) = 5.92$, MSE = 0.044, and on D', $F(2, 66) = 7.85$, MSE = 0.299. The decrease in intercept with category length was non-monotonic, with the worst discrimination for category length 8 (Figure 4c) and the best for category length 4 (Figure 4b). For D', in contrast, the decrease was monotonic and decelerating with category length.

Performance for dissimilar and similar words was compared using a two-way (similarity x category length) ANOVA excluding data from the zero category length condition. The effect of similarity on slope was highly significant, $F(1, 33) = 27.3$, MSE = 0.044, but the interaction between similarity and category length was not significant, $F(2, 66) = 1.23$, MSE = 0.037. Similarity also had a highly significant effect on both the intercept, $F(1, 33) = 49.5$, MSE = 0.107, and D', $F(1, 33) = 87.7$,

MSE = 0.449, with dissimilar accuracy being greater than similar accuracy in both cases. The interaction between similarity and category length was marginally significant for the intercept, $F(2, 66) = 2.89$, MSE = 0.062, $p = 0.063$, and highly significant for D', $F(2, 66) = 9.25$, MSE = 0.432, reflecting an increase in D' with category length for dissimilar words and a decrease with category length for similar words.

In order to facilitate comparison with the results of Shiffrin et al. (1995), Table 1 reports average Z(H) and Z(FA) values and corresponding H and FA probabilities, for binary new-old decisions. Similarity caused a large increase in false alarms and a smaller increase in hits for all category lengths. The same pattern was obtained for Experiments 1 and 2. In agreement with Shiffrin et al.'s results, false alarms increased with category length, but hits were relatively constant.

Overall, the category length results for similar items in Experiment 3 conform to the predictions made by SAM: decreased accuracy and increased slope with increased category length, although the increase in slope was only evident for the longest category. SAM's predictions for dissimilar slope were confirmed, with dissimilar slope being less than similar slope and unaffected by category length. However, there was some evidence that accuracy for dissimilar items increased with category length, whereas SAM predicts that dissimilar accuracy is unaffected by category length, and hits were not much affected by category length.

Shiffrin et al. (1995) explained the lack of increase in hits with category length by suggesting that increased category length may decrease recollection. DPS theory predicts that as recollection decreases ZROC functions become more linear and slope approaches one. Although similar slope did increase with category length, no upward concavity was evident for the shorter category lengths. It is possible that even the

shortest category in Experiment 3 was long enough to discourage any use of recollection. However, if this were true, DPS theory cannot explain the change in slope with category length, because only changes in recollection can cause changes in slope.

Glanzer et al. (1993) showed that Attention Likelihood Theory predicts lower slopes for conditions that have higher accuracy due to greater attention at study. This prediction was not borne out for the effect of orthographic similarity in the present experiment and Experiment 2. Accuracy was less for similar than dissimilar words, but similar slope was greater than dissimilar slope. Decreased attention to similar words also predicts a mirror effect in hits and false alarms. A mirror effect occurs when a less accurate condition (e.g., similar words) produces more false alarms but less hits than a more accurate condition (e.g., dissimilar words). However, as shown in Table 1, similar words produced more false alarms, but also more hits, than dissimilar words, and the same pattern was found for Experiments 1 and 2 (see also Shiffrin et al., 1995). These results indicate that Attention Likelihood Theory cannot explain similarity effects, at least through differential study attention.

Experiment 4

The results of the similarity manipulations in Experiments 1-3 are broadly in agreement with item noise theories (Dennis & Humphreys, 2001), which postulate that familiarity is based on the combined matches of a test probe to memory images of the study list items. Study list items that are similar to the probe produce larger match values, and so increase the familiarity of both new and old similar probes. However, Dennis and Humphreys suggested that item similarity effects might be caused by criterion shifts or uncontrolled differences between similar and dissimilar words, rather than by item noise.

Although the ZROC analyses of Experiments 1-3 indicate that similarity effects cannot be explained by criterion shifts alone, they may have been produced by uncontrolled differences between similar and dissimilar words, because separate sets of words were used to make up the similar and dissimilar items in each list. Dissimilar words were selected so that they had low similarity with other dissimilar words and with the similar words, and so that similar and dissimilar words were matched on word frequency. However, similar and dissimilar words may have differed on other characteristics.

Experiment 4 was designed to check the effect of uncontrolled differences between the orthographically similar and dissimilar word pools. For each of the four word-lengths (5-8 letters), study lists were constructed that contained one word from each of the similar word pools. Participants completed five blocks, with each block consisting of a set of four study-test cycles, so that within each block only one word from each similar pool was presented. The aim of the design was to maximise the spacing between presentations of words from the same similar word pool, so that participants would be unlikely to use different decision criteria for dissimilar and similar word-pool items. As a result, the NUS model could be fit simultaneously to dissimilar and similar word data. Due to constraints on the number of stimuli, list length was reduced to 24 in Experiment 4. Each study pair was presented for two seconds instead of the three seconds used in earlier experiments in order to avoid ceiling effects.

Methods

Participants

The 22 participants were students at the University of Newcastle, Australia who were paid $10 for attending a one-hour session.

Apparatus and Stimuli and Procedure

The dissimilar words used in Experiment 4 were drawn from the 32 dissimilar words in each of the 36 mixed similar and dissimilar word sets used in Experiment 3. Hence, 288 five-letter words, 384 six-letter words, 288 seven-letter words and 192 eight-letter dissimilar words were used to construct study and test lists. The similar words were drawn from the 24 similar words in each of the 34 mixed similar and dissimilar word sets used in Experiment 3. Two similar word sets used in Experiment 3 were excluded (one five-letter list and one seven-letter set with column headings PRI_E and S__PPER in the Appendix) so that there were even numbers of similar words sets at each word length: 8 five-letter sets, 12 six-letter sets, 8 seven-letter sets, and 6 eight-letter sets. Words in each study-test list had the same number of letters.

Lists were constructed by selecting one similar word randomly without replacement from each of the sets containing words with the same number of letters. Hence, the number of similar words used to construct a list depended on the word length in a list, 8 for five-letter, 12 for six-letter, 8 for seven-letter, and 6 for eight-letter lists. Half of the similar words were randomly selected for study, so there were equal numbers of new and old similar words for each study-test list. The remaining words required to construct a study-test list were drawn randomly from the appropriate word-length pool of dissimilar words. Study lists consisted of twelve pairs of words presented for two seconds each. The first two pairs and last two pairs were drawn from the dissimilar pool. Only one of these "buffer" words was selected randomly for testing, and the results for it were not recorded. The middle eight study-list pairs were constructed from both similar and dissimilar words, with order and pairing determined randomly. All of these 16 old words, and equal numbers of similar and dissimilar new words, were tested.

Study-test lists were arranged in blocks of four, consisting of one list of each word length presented in random order. Hence, within each block only one word from each different similar-word set was used. Each block produced results for 17 old and 17 new similar words, and 47 old and 47 new dissimilar words. Participants first performed a practice study-test cycle, as in earlier experiments, using a list made entirely from dissimilar words. They then completed five blocks, for a total of 20 study-test cycles. Procedures were otherwise the same as in earlier experiments.

Results and Discussion

Because there were only 17 tests of similar words per block, the NUS model could not be fit to the confidence data from each block separately. Instead, data were aggregated over the first two blocks and the last two blocks before fitting. Analysis of the first two blocks allowed the effect of word type (similar or dissimilar) to be determined relatively free from the effect of studying other similar words from the same set. Comparison of the results for the first two and last two blocks allowed assessment of the effect of four to five presentations of similar words from the same set in different blocks. Analyses aggregating data from all blocks, and aggregating data from the first three blocks, produced corresponding results, and so are not reported here.

Figure 5 illustrates the results for fits that assumed the same confidence criteria for similar and dissimilar word data, but different criteria for the first and last two blocks, with the new dissimilar word conditions used as a reference. Visually, the fits for the first two blocks (Figure 5a) were not as good as in the previous experiments, but only 2 of the 22 participants had significant misfit at the 0.01 level according to the χ^2 test. Fit was much better for the last two blocks (Figure 5b), with no participants having significant misfit at the 0.01 level. As in earlier experiments,

the pattern of deviation was systematically concave downward, consistent with some guessing in the first two blocks that largely disappeared by the last two blocks of the experiment[4].

The only significant difference between similar and dissimilar word intercepts occurred for new probes in the last two blocks, $F(1, 21) = 6.58$, MSE = 0.021, with similar new words having a higher intercept than dissimilar new words. For both the first and last two blocks, similar slope was greater than dissimilar slope for both new probes (first two blocks: $F(1, 21) = 22.8$, MSE = 0.02, last two blocks: $F(1, 21) = 13.6$, MSE = 0.027) and old probes (first two blocks: $F(1, 21) = 8.16$, MSE = 0.008, last two blocks: $F(1, 21) = 4.95$, MSE = 0.011). The difference between similar and dissimilar slopes was greater for new than old probes, as confirmed by significant interactions for both the first two blocks, $F(1, 21) = 6.35$, MSE = 0.014, and last two blocks, $F(1, 21) = 5.6$, MSE = 0.013.

Overall, Experiment 4 showed some weak but reliable differences between dissimilar and similar word pools, particularly in ZROC slope. However, these differences were in the opposite direction to the differences observed in earlier experiments. The slopes reported for the similar conditions in Experiments 2 and 3 estimate the standard deviation of old similar relative to new similar, whereas in the present experiment the old similar slope estimates are relative to the new dissimilar condition. In order to compare the old similar slopes from Experiment 4 to the similar slopes in earlier experiments, the old similar slope estimates in Figure 5 must be divided by the corresponding new similar slope estimates. The resulting values are 0.70 for the first two blocks and 0.76 for the last two blocks. Both of these values are less than the corresponding dissimilar old slope estimates for Experiment 4, by an average of 0.07. For Experiments 2 and 3, in contrast, the slopes for similar old words

were on average greater than the slopes for dissimilar old, by 0.23 and 0.15 respectively. Hence, differences between the similar and dissimilar word pools could only have decreased the similarity effects on slope seen in Experiments 2 and 3.

Experiment 4 also demonstrated that inter-list proactive interference effects were relatively weak compared to the strong intra-list interference effects found in earlier experiments. If similarity effects in earlier experiments accrued over lists, an interaction between blocks and the difference between similar and dissimilar conditions would be expected in the present experiment, but none was found. Instead, performance for similar and dissimilar words decreased equally over blocks. The relatively weak inter-list proactive interference found in Experiment 4 is also inconsistent with Hirshman and Hostetter's (2000) suggestion that slope effects are swamped by large inter-list proactive interference effects in designs using many study lists.

The results of Experiment 4 indicate that similarity effects found in earlier experiments were not caused by uncontrolled differences between the orthographically similar and dissimilar word pools. However, these results do not necessarily reject Dennis and Humphrey's (2001) context noise model, which does not predict item noise effects. First, as suggested by Dennis and Humphreys, when participants are aware of the categorised nature of the lists, similarity effects may be caused by implicit associative responses during study. Second, orthographically similar words tended to share affixes, which may have led participants to encode them by components (e.g. affix + word body). In Dennis and Humphrey's theory, the corresponding item representation would be distributed, and so item noise effects could result.

General Discussion

Theories of recognition memory often rely on changes in the distribution of familiarity to explain important phenomena such as the constancy-of-slopes generalization (Yonelinas, 1994), the list-length effect (e.g., Gronlund & Elam, 1994), the null list-strength effect (e.g., Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Ratcliff, et al., 1994), similarity and category length effects (e.g., Hintzman, 1986, 2001; Shiffrin, Huber, & Marinelli, 1995) and the mirror effect (e.g., Glanzer, Adams, & Iverson, 1991). Hence, the measurement of the distribution of familiarity by ROC analysis provides an important constraint for theories. The experiments reported here tested two accounts of ROC results, Ratcliff et al.'s (1994) constancy-of-slopes generalization, and Yonelinas' (1994) Dual-Process Signal-Detection (DPS) Theory.

Dual-Process Signal-Detection Theory

A clear result from the present experiments is that the DPS theory of Yonelinas (1994) does not provide a general explanation of ROC results in item recognition. DPS theory can only explain ZROC slopes less than one by concave upward ZROCs. Slopes much less than one were observed in almost all conditions of the present experiments, but ZROCs were not concave upward. These results agree with the findings of Glanzer et al. (1999a) and extend them. Glanzer et al. allowed that concave upward ZROCs might be found with short lists and powerful encoding manipulations. The present experiments fulfilled both conditions and still no evidence of concave upward ZROCs was found. Similarly, Westerman (2001) suggested that: "One recognition situation that seems to call for recollection rather than a general assessment of familiarity occurs when participants are asked to discriminate between targets and lures that are very similar to each other" (p. 723). Although very similar

targets and lures were used in the present experiments, there was no evidence for recollection as modelled by DPS.

Yonelinas's (1999a) least-squares-probability fitting method was applied to the individual participant data from the first three experiments of the present paper. Each condition was fit separately using only cases with non-zero hit and false alarm probabilities for at least three points, as a minimum of three points is required to differentiate the models. Although average R^2 values differed in only the fourth decimal place, the NUS model consistently provided a better fit than DPS theory, winning 75.5% of cases in Experiment 1, 77.8% of cases in Experiment 2, and 78.6% of cases in Experiment 3. Overall, less than 10% of subjects had a majority of conditions with superior DPS fits, and no condition in any experiment ever had a majority of subjects with superior DPS fits. As was the case for the data from Glanzer et al. (1999a) taken from Yonelinas's Figure 3, binomial tests significantly favour the NUS model in every experiment and experimental condition.

These findings strongly reject Yonelinas's (1994) DPS theory as a model of item recognition ROCs, and as an explanation of findings about ZROC slope in item recognition. They do not, however, rule out some other type of dual-process theory, or a role for recollection in item recognition. DPS theory could accommodate linear ZROCs with slopes less than one if its signal detection component allowed unequal new and old familiarity variance. With this extension, however, the results from the present experiments indicate that the recollection component of DPS plays little role in item recognition despite conditions that should have promoted recollection (short lists, good encoding and similar targets and lures). The extended DPS model also lacks a principled explanation of the large changes in ZROC slope seen in the present experiments.

In common with some other dual process theories (e.g., Jacoby, 1991), DPS assumes that recollection is an all-or-none process. When memory relies on all-or-none recollection and guessing, ROCs, rather than ZROCs, will be linear (Hilford et al., 2002). However, other theories, such as Johnson, Hashtroudi and Lindsay's (1993) source monitoring framework, assume that recollection can provide continuous information (see e.g., Dodson, Holland & Shimamura, 1998, for experimental evidence). When continuous, normally distributed sources of information are combined linearly the NUS model will apply and linear ZROCs will be found. A quantitative example of this approach is Bank's (2000) multivariate signal detection model of recognition memory. For such models, the linear ZROCs found in the present experiments could be consistent with a role for recollection in item recognition.

DPS theory has also been applied to recognition memory paradigms that might be argued to rely more heavily on recollection than item recognition, such as source and associative recognition. Linear ROCs were found in associative recognition by Yonelinas (1997) and Yonelinas et al. (1999), in source recognition by Yonelinas (1999b) and in item recognition where a source judgement was also required by Yonelinas et al. (1996). However, Qin et al. (2001) and Hilford et al. (2002) found curvilinear ROCs and close to linear ZROCs in source recognition. Qin et al. suggested that source ROCs are usually curvilinear, with linear ROCs only occurring when source information is very impoverished. Hilford et al. suggested that source information is continuous, but that participants sometimes fail to encode source information. In associative recognition, Kelly and Wixted (2001) found that ROCs changed from close to linear in weak study conditions to curvilinear (and close to ZROC linear) for stronger study conditions. On the basis of their results Kelly and

Wixted suggested that associative information is "some-or-none", meaning that associative information is sometimes not available, but when it is available it is continuous. Both Hilford et al. and Kelly and Wixted proposed mixture models (see also DeCarlo, 2002) able to provide good accounts of the deviations from ZROC linearity on the basis of variations in the probability of source encoding and the availability of associative information respectively.

Further consideration of source and associative recognition is beyond the scope of the present paper. However, it can be concluded that these paradigms provide stronger evidence for a contribution from recollection (e.g., recall-to-reject, Rotello et al., 2000) and stronger evidence of deviations from ZROC linearity than item recognition paradigms. Although DPS does allow a role for recollection, its account of ROC shape for these paradigms is incomplete. Alternative approaches, which assume that recollection provides continuous information, perhaps augmented with a mixture mechanism to account for failures to encode or retrieve details, appear more likely to provide a general model of recognition memory.

The Normal Unequal-Variance Signal-Detection Model

In contrast to DPS theory, the NUS model provided an excellent description of the item recognition ROCs in the present experiments. Although the present results validate the general application of the Theory of Signal Detection (Green & Swets, 1966), and the NUS model in particular, to item recognition memory performance, evidence was found for a small but consistent excess of high confidence errors. The resulting small concave downward deviation found in ZROCs is consistent with Ratcliff et al.'s (1994) guessing model but does not uniquely support it. It is important to note, however, that these deviations were very small; the largest deviations in each experiment were on average around 0.005 in terms of probability or an excess of one high confidence error for one third of the participants.

Van Zandt (2000) provided a fundamental challenge to the NUS model of recognition memory confidence by testing its prediction that ZROC parameters are invariant under changes in bias. For bias manipulations using both prior odds and payoffs, she found that ZROC slope increased with increasing old bias. These findings cast doubt on the assumption that confidence judgements are scaled directly from familiarity, or from posterior odds values derived from familiarity (e.g., Glanzer et al., 1991; Shiffrin & Steyvers, 1997). In the present experiments, participants may have developed a bias to respond new for similar words in order to counteract high false alarm rates. Van Zandt's results suggest that a bias to respond new causes a reduction of ZROC slope. However, ZROC slope for similar words was significantly greater than ZROC slope for dissimilar words. Hence, changes in slope due to changes in bias cannot explain the larger slopes for similar than dissimilar words, although such changes may have reduced the magnitude of the difference.

Van Zandt (2000) also developed a Poisson race model, based on the balance of evidence hypothesis of Vickers (1979), which was able to account for changes in ZROC slope with bias, and for response time in making confidence decisions. An account of response times is important in order to rule out possible speed-accuracy tradeoff. Although the present experiments were not explicitly designed to investigate response time, some indicative analyses[5] can be performed on results from Experiments 1 to 3. ANOVAs examining the mean time for correct responses, and using confidence and new versus old response factors, as well as stimulus factors such as similar versus dissimilar words, weak versus strong study, and category length, produced a consistent pattern of results for all experiments: response factors dominated response time effects. On average across Experiments 1 to 3, the main effect of response confidence accounted for 84% of the systematic variance and new

versus old responses accounted for 10% of the systematic variance. New responses were slower than old responses, and response time decreased markedly with confidence ($p < .001$ in all cases). In contrast, stimulus factors rarely produced significant effects[6].

These results indicate that the response methodology used here, simultaneously providing both response confidence and the new versus old decision, decoupled response confidence probabilities from response times, at least to a first approximation. They also indicate that the ZROC analyses of stimulus factors reported above were not confounded by speed-accuracy trade-off. That is, given that two stimuli from different conditions were given the same confidence rating, they did not differ in response time. The final section considers the theoretical implications of stimulus factor effects on ZROC parameter estimates found in the present experiments.

ZROC Parameters

The parameter estimates for the NUS model in the present experiments did not support the constancy-of-slopes generalization of Ratcliff et al. (1994), with the exception of massed study effects. Although proving a null hypothesis is difficult, it is important to note that null hypotheses can, and should, be confirmed where a "good effort" (Frick, 1995) has been made to find an effect. In the present experiments massed study had a weak and inconsistent effect on slope. These results support the findings of Ratcliff et al. (1992) and Ratcliff et al. (1994) that massed study provides a case where a null effect on ZROC slope can reasonably be accepted. As previously argued, the apparently contradictory findings of Glanzer et al. (1999a) that massed study affected slope were likely due to the inclusion of low accuracy conditions.

The strongest violations of the constancy-of-slopes generalization were due to orthographic similarity and to a lesser degree semantic similarity. These results are

consistent with Ratcliff et al.'s (1994) marginally significant similarity effects on ZROC slope, and indicate that their failure to find a significant effect may have been due to a lack of power and weak effects due to the use of short categories. Constancy-of-slopes was also violated by the results for category length, but only for the longest category (1/2 of the study list) compared relative to the two shorter categories ($1/6^{th}$ and $1/3^{rd}$ of the study list). In agreement with previous findings (e.g., Gillund & Shiffrin, 1984) the effect of orthographic similarity on accuracy was much more powerful than the effect of semantic similarity. Semantic similarity did not produce a consistent accuracy deficit in the present experiments, despite the strength of the semantic manipulation. However, semantic similarity did have significant effects on ZROC slope.

ZROC parameter estimates produced two surprising findings from the viewpoint of the global memory models. First, accuracy increased with category length for dissimilar words. The global memory models predict constant accuracy, because the match between a dissimilar probe and the study list is unaffected by category length. A possible explanation is that participants were able to exclude similar word memory traces from the matching process, perhaps by using an appropriate context cue, so that the effective list length for dissimilar words decreased with increasing category length. However, for SAM and MINERVA2, this mechanism predicts that dissimilar ZROC slope should decrease with category length, but a decrease was not observed.

The second surprising finding was an increase in ZROC slope caused by spaced study. The effect of spaced study on slope was not as great as the effect of similarity, but it was consistent in Experiments 1 and 2 for both dissimilar and similar words. This result implies that changes in familiarity variance with study can be non-

monotonic, with variance increasing due to the first study episode, and then decreasing with further spaced study episodes. Although the increase in slope is surprising for the global memory models, which predict that slope either decreases or remains constant, it may help to explain the apparently contradictory results of Glanzer et al. (1999a) and Ratcliff et al. (1994) for spaced study. Depending on which pairs of points on the non-monotonic relationship between accuracy and slope are compared, ZROC slope may decrease, remain constant, or increase.

Decreased familiarity variance with spaced study may partially explain the accuracy advantage of spaced over massed study in item recognition. However, variance differences cannot explain all of the advantage of spaced over massed study. Fortuitously, the new dissimilar distributions for massed and spaced conditions were almost identical[7] in Experiments 1 and 2. Equality of the new distributions allows the massed and spaced dissimilar $\underline{D}$'s to be compared directly as measures of changes in mean familiarity. As shown in Figures 2 and 3, $\underline{D}'$ was larger for strong spaced than strong massed study in both Experiments 1 and 2, indicating spaced study causes a greater increase in mean familiarity than massed study. Hence, the spaced study advantage appears to be subserved by changes in both the mean and variance of familiarity.

An extension of Greene's (1989) and Challis's (1993) deficient processing theories may be able to explain the ZROC results for massed and spaced study. Both theories suggest that the effectiveness of study decreases as items receive more study. Mean familiarity is greater for spaced study than massed study because the effects of prior study decrease, and hence the effectiveness of study increases, as the space between repetitions increases. The same mechanism can also account for effects on familiarity variance. If an item receives unusually effective study on the first

presentation, study will be less effective on the second presentation. Conversely, if study is unusually ineffective on the first presentation, study for the second presentation will tend to be more effective. The net result is a negative correlation between the familiarity generated by the first and second study episodes. If the negative correlation is sufficiently strong, studying two spaced presentations may result in a reduction in familiarity variance relative to a studying one presentation.

A similar mechanism may explain the increase in ZROC slope with category length seen in Experiment 3, if similar words act like partial repetitions. However, there is evidence that the familiarities of similar items tend to be positively correlated (Hintzman, 2001), which would tend to increase familiarity variance as category length increases. It is possible that cancellation of these two opposite tendencies might explain the invariance of ZROC slope for the two shortest categories in Experiment 3. Further work is needed to test these possibilities.

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: Volume 1, Learning, memory and thinking (pp. 242-293). San Francisco: Freeman.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. Psychological Science, 11, 267-273.

Casey, P. J. (1988). Category norms for Australians. Australian Journal of Psychology, 40, 323-339.

Challis, B. H. (1993). Spacing effects on cured-memory tests depend on level of processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19, 389-396.

Coltheart, M. (1981). The MRC psycholinguistic database, Quarterly Journal of Experimental Psychology, 33A, 497-505.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. Psychological Review, 109, 710-721.

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

Dodson, C. S., Holland, P. W. & Shimamura, A. P. (1998). On the recollection of specific- and partial-source information. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 1121-1136.

Donaldson, W., & Murdock, B. B., Jr. (1968). Criterion change in continuous recognition memory. Journal of Experimental Psychology, 76, 325-330.

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.

Frick, R. W. (1995). Accepting the null hypothesis. Memory & Cognition, 23, 132-138.

Gehring, R. E., Toglia, M. P., & Kimble, G. A. (1976). Recognition memory for words and pictures at short and long intervals. Memory & Cognition, 4, 256-260.

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

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

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

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

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

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. Journal of Verbal Learning and Verbal Behavior, 15, 1-16.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics, New Your: Robert E. Kreiger Publishing.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 371-377.

Grey, D. R., & Morgan, B., J., T. (1972). Some aspects of ROC curve-fitting: Normal and logistic models. Journal of Mathematical Psychology, 9, 128-139.

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.

Heathcote, A. (1988). Screen control and timing routines for the IBM microcomputer family using a high-level language. Behaviour Research Methods, Instruments & Computers, 20, 289-297.

Hilford, A., Glanzer, M., Kim, K. & DeCarlo, L. (2002). Regularities of source recognition: ROC analysis. Journal of Experimental Psychology: General, 131, 494-510.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. Psychological Review, 93, 411-428.

Hintzman, D. L. (2001). Similarity, global matching, and judgments of frequency. Memory & Cognition, 29, 547-556.

Hirshman, E. & Hostetter, M. (2000). Using ROC curves to test models of recognition memory: The relationship between presentation duration and slope. Memory & Cognition, 28, 161-166.

Humphreys, M. S., Pike, R., Bain, J. D. & Tehan, G. (1989). Global matching: A comparison of SAM, Minerva II, Matrix and TODAM models. Journal of Mathematical Psychology, 33, 36-67.

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

Johnson, M. L., Hashtroudi, S. & Lindsay, D. S. (1993). Source monitoring. Psychological Bulletin, 114, 3-28.

Keeley, 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.

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.

Mandler, G. (1980). Recognizing: The judgement 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.

Murdock, B. B. (1998). The mirror effect and attention-likelihood theory: A reflective analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 524-534.

Murdock, B. B., & Kahana, M. J., (1993). An analysis of the list-strength effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19, 689-697.

Murnane, K. & Shiffrin, R. M. (1991). Word repetition in sentence recognition. Memory & Cognition, 19, 119-130.

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

Qin, J., Raye, C. L., Johnson, M. K. & Mitchell, K. J. (2001). Source ROCs are (typically) curvilinear: Comment on Yonelinas (1999). Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1110-1115.

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

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

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

Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. Journal of Experimental Psychology: General, 124, 352-374.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107, 358-367.

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., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 267-287.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 179-195.

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

Van Zandt, T. (2000). ROC Curves and Confidence Judgments in Recognition Memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 582-600.

Vickers, D. (1979). Decision processes in visual perception. New York: Academic Press.

Westerman, D. L. (2001). The role of familiarity in item recognition, associative recognition and plurality recognition on self-paced and speeded tests. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 723-732.

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

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

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

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. Consciousness and Cognition, 5, 418-441.

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I. G. & Soltani, M. (1999). Recognition memory of faces: When familiarity supports associative recognition judgements. Psychonomic Bulletin & Review, 6, 654-661.

Footnotes

[1] http://archives.math.utk.edu/software/mac/miscellaneous/DataThief/, Version 2.0, 1994, by Kees Huyser and Jan van der Laan.

[2] Note that differences between the models are hard to see in probability plots as in Yonelinas's (1999a) Figure 3 but are clearer in Z transformed plots. When R^2 was calculated on observed and predicted Z values rather than probabilities, the average difference in fit between the models was also larger, occurring in the third decimal place.

[3] Empirical corrections often used to obtain defined Z estimates when probabilities are zero or one are not appropriate for ZROC estimation where accuracy is high because cases occur were intermediate as well as extreme categories have undefined Z values (e.g. no false alarms in the two most extreme old response confidence categories).

[4] A reviewer suggested that the excess of high confidence errors in Experiments 1-3 might have also been confined to early study-test cycles. However, an analysis of high confidence errors from these experiments did not reveal a consistent pattern. High confidence errors showed minor increases or decreases, or remained constant, as a function of study-test cycles, with no apparent consistency.

[5] Response time was measured to millisecond accuracy but responses were made through the computer keyboard, which is a buffered device that introduces uniform noise of approximately ± 20 ms. Participants always used their left hand for new responses and their right hand for old responses, so comparisons of new and old response times are confounded by the effects of handedness. ANOVAs using response factors such as new versus old responses and confidence violate the fixed effects and independence assumptions of ANOVA. To partially compensate for these violations, a conservative significance criterion of 0.01 was adopted, and Greenhouse-Geisser corrections were applied to degrees of freedom. ANOVAs involving response confidence were restricted to correct responses because of missing values for incorrect responses. Even for correct responses, some participants still had missing values and had to be excluded. For the massed and spaced conditions of Experiment 1 two and three participants, and for Experiment 2 one and two participants, respectively, were excluded. For Experiment 3, 7 participants were excluded. Experiment 4 was not examined because its lower number of observations resulted in to many missing values to make analysis viable.

[6] Out of the 44 tests involving stimulus factors, the only exceptions were in the massed condition of Experiment 1, where similarity and new versus old responses interacted, $F(1, 29) = 11.7$, MSE = 132154, the spaced condition of Experiment 1, where the three-way interaction between similarity, new versus old response and confidence interacted, $F(2, 56) = 5.32$, MSE = 87413, and Experiment 3, where similarity and confidence interacted, $F(1.9, 48.6) = 7.07$, MSE = 83608.

[7] Neither false alarm probabilities or Z transformed false alarm probabilities differed significantly between dissimilar-new massed and spaced study conditions in Experiment 1 ($F < 1$ for false alarms and $F(1, 62) = 1.61$, MSE = 0.560 for Z transformed false alarms) or Experiment 2 (both $Fs < 1$).

Author Note

Thanks to two anonymous reviewers, Scott Brown, Kerry Chalmers, Beth Johns, Richard Heath, Bill Hockley, Ching-Fan Sheu, and Doug Mewhort for suggestions to improve readability. This research was supported by grants from the Research Management Committee and from the Faculty of Science, University of Newcastle, Australia.

Appendix: Word Sets

Table A1. Semantically similar word lists used in Experiment 1 and corresponding category names.

ANIMAL	BIRD	BODY PART	CLOTHING	COLOUR	CRIME
ANTELOPE	CANARY	ARM	BELT	BEIGE	ABDUCTION
APE	CHICKEN	BOTTOM	BLAZER	BLACK	ADULTERY
BULL	CROW	BRAIN	BLOUSE	BLUE	ARSON
CAMEL	DOVE	BREAST	BOOTS	BONE	ASSAULT
CAT	DUCK	CHEEK	CAP	BRONZE	BATTERY
COW	EAGLE	CHIN	COAT	BROWN	BLACKMAIL
DEER	FALCON	EAR	DRESS	CREAM	BURGLARY
ELEPHANT	GOOSE	ELBOW	HAT	CRIMSON	EXTORTION
FOX	HAWK	FACE	JACKET	GREEN	FORGERY
GOAT	HEN	FINGER	JUMPER	GREY	FRAUD
GORILLA	OWL	FOOT	OVERCOAT	LILAC	INCEST
LEOPARD	PARROT	HAIR	PANTS	MAROON	KILLING
LION	PEACOCK	KNEE	PETTICOAT	NAVY	LARCENY
LIZARD	PHEASANT	LEG	SCARF	OLIVE	LYING
MONKEY	PIGEON	LIP	SHIRT	PINK	MANSLAUGHTER
MOUSE	QUAIL	MOUTH	SKIRT	PURPLE	MURDER
PIG	ROBIN	NECK	SPENCER	RED	PERJURY
RABBIT	ROOSTER	NOSE	SUIT	ROSE	PROSTITUTION
RAT	SPARROW	SHOULDER	SWEATER	RUST	RAPE
RHINO	SWALLOW	SKIN	TIE	SCARLET	ROBBERY
SEAL	SWAN	STOMACH	JEANS	MAUVE	SLANDER
SHEEP	TURKEY	TEETH	TROUSERS	TAN	SUICIDE
SNAKE	VULTURE	THROAT	UNDERWEAR	VIOLET	THEFT
TIGER	WREN	WRIST	VEST	YELLOW	TREASON

DISEASE	DRINK	DRUG	EMOTION	FRUIT	FURNITURE
ALCOHOLISM	BOURBON	ACID	ANGER	APPLE	ARMCHAIR
APPENDICITIS	BRANDY	ALCOHOL	ANXIETY	APRICOT	BAR
ARTHRITIS	CHAMPAGNE	ASPIRIN	CONFUSION	AVOCADO	BED
ASTHMA	CIDER	BARBITURATE	CRY	BERRY	BENCH
BRONCHITIS	COCOA	BEER	DEPRESSION	BLACKBERRY	CABINET
CANCER	COFFEE	CIGARETTE	DESPAIR	CHERRY	CARPET
CHOLERA	COKE	COCAINE	EXCITEMENT	COCONUT	CHAIR
COLD	CORDIAL	DOPE	FEAR	DATE	CHEST
DIABETES	GIN	GLUE	GRIEF	FIG	COUCH
EMPHYSEMA	JUICE	GRASS	GUILT	GRAPE	DESK
FLU	LEMON	HASH	HAPPINESS	MELON	DRAWER
GANGRENE	LEMONADE	HEMP	HOPE	MULBERRY	LAMP
HEART	LIME	HEROIN	HURT	NUT	LIGHT
HEPATITIS	MARTINI	INSULIN	JOY	PEACH	LOUNGE
INFLUENZA	MILK	MEDICINE	LAUGHTER	PEAR	MIRROR
LEPROSY	ORANGE	MORPHINE	MAD	PERSIMMON	PIANO
MALARIA	PORT	NICOTINE	PAIN	PINEAPPLE	PICTURE
MEASLES	RUM	OPIUM	PASSION	PLUM	RADIO
MUMPS	SCOTCH	PENICILLIN	PEACE	POMEGRANATE	SHELF
PNEUMONIA	SHERRY	POT	RELIEF	PRUNE	SINK
SMALLPOX	SODA	SMACK	SCREAM	QUINCE	SOFA
SYPHILIS	SQUASH	SPEED	SORROW	RASPBERRY	STOOL
TUBERCULOSIS	WATER	TEA	SURPRISE	STRAWBERRY	STOVE
TYPHOID	WHISKY	TOBACCO	SYMPATHY	TOMATO	WARDROBE

Table A1 continues

MINERAL	OCCUPATION	TOY	TREE	VEHICLE	WEAPON
AMBER	ARCHITECT	AEROPLANE	ACACIA	BUGGY	ARROW
COAL	ARTIST	BALL	ASH	BUS	BATON
COPPER	BAKER	BAT	BEECH	CAR	BOMB
CORAL	BUILDER	BEAR	BIRCH	CARAVAN	BOW
CRYSTAL	BUTCHER	BICYCLE	BOX	CARRIAGE	BRICK
DIAMOND	CLEANER	BOAT	CEDAR	CART	CHAIN
EMERALD	CLERK	BOOK	CHESTNUT	FERRY	CLUB
GEM	COOK	DOG	CHRISTMAS	FORD	FEET
GLASS	DENTIST	DOLL	CYPRESS	JEEP	FIST
GOLD	DRIVER	DRUM	ELM	LORRY	GAS
GRANITE	ENGINEER	FOOTBALL	EUCALYPTUS	PLANE	HAMMER
IRON	FARMER	GAME	FERN	SEDAN	KNIFE
JADE	LAWYER	GUN	FIR	SHIP	PISTOL
JASMINE	MANAGER	HORSE	GUM	TANK	RAZOR
JET	NURSE	KITE	MAPLE	TAXI	REVOLVER
JEWEL	PAINTER	MODEL	OAK	TRACTOR	RIFLE
MARBLE	PILOT	MONOPOLY	PALM	TRAILER	ROCK
NICKEL	POLICEMAN	PRAM	PEPPERCORN	TRAIN	ROCKET
ONYX	POLITICIAN	PUZZLE	PINE	TRAM	ROPE
PEARL	SALESMAN	RATTLE	POPLAR	TRUCK	STICK
RUBY	SCIENTIST	SOLDIER	REDWOOD	UTILITY	STONE
SAPPHIRE	SECRETARY	SWING	SYCAMORE	VALIANT	SWORD
SILVER	STUDENT	TOP	WALNUT	VAN	BAYONET
TURQUOISE	TEACHER	TRICYCLE	WILLOW	WAGON	WHIP

Table A2. Orthographically similar word lists used in Experiments 2 and 3, with column headings indicating letters that appear in more than 50% of words.

S_A_E	_RA_E	_O_ER	__RRY	S_EE_	__TCH	__ING	___CK	PRI_E	___ENT	__GGED
SHARE	GRAVE	LOVER	HARRY	STEER	PATCH	SLING	SLICK	PRIDE	PATENT	LOGGED
STARE	BRAVE	LOWER	PARRY	SHEER	PITCH	SWING	SLACK	PRIME	RECENT	RAGGED
SHAKE	GRACE	BOWER	TARRY	STEEP	WATCH	CLING	CLICK	PRICE	ASCENT	DOGGED
SHAVE	CRAVE	LOSER	CARRY	STEED	BATCH	STING	STICK	PRIZE	RESENT	BAGGED
SPARE	GRATE	HOVER	BARRY	STEEL	HATCH	FLING	FLICK	TRIBE	LATENT	JAGGED
STATE	GRAZE	ROVER	MARRY	SHEEP	MATCH	SUING	SHACK	BRIBE	ACCENT	LEGGED
SHADE	GRADE	POWER	PERRY	SNEER	CATCH	TYING	STACK	URINE	ASSENT	RUGGED
STAKE	GRAPE	DOWER	TERRY	SHEET	LATCH	DYING	BLACK	TRITE	POTENT	BEGGED
SNARE	BRAKE	TOWER	MERRY	SHEEN	BITCH	LYING	SHOCK	BRIDE	ARDENT	LOGGER
SCARE	BRACE	COVER	BERRY	SWEEP	DITCH	OWING	STOCK	DRIVE	REPENT	RIGGER
SHAME	CRATE	BORER	FERRY	SLEEP	WITCH	DOING	CLOCK	ARISE	ABSENT	DIGGER
SHAPE	CRAZE	BOXER	HURRY	SWEET	HITCH	BRING	BLOCK	TRIPE	LAMENT	RINGED
SPATE	TRACE	ROGER	CURRY	SHEAR	NOTCH	GOING	FLOCK	WRITE	DECENT	BIGGER
STALE	CRANE	SOBER	SORRY	SLEET	DUTCH	BEING	SMACK	PRONE	CEMENT	LUNGED
SLATE	DRAKE	HOMER	WORRY	SPEED	FETCH	THING	STUCK	CRIME	RODENT	JIGGER
SPADE	TRADE	POKER	HARDY	SLEEK	PINCH	SLUNG	CLUCK	PROBE	PARENT	HINGED
STAGE	IRATE	FOYER	HARPY	SWEAR	PORCH	WRING	SNACK	PROVE	ADVENT	WINGED
SNAKE	FRAME	VOTER	HAIRY	STEAD	PERCH	ICING	CHICK	PROSE	INTENT	HANGED
SKATE	ERASE	LEVER	PARTY	STEAL	POACH	SLANG	FLECK	PRUNE	TALENT	LODGED
SPACE	GROVE	LIVER	TARDY	STEAK	POUCH	SWUNG	PLUCK	PRINT	MOMENT	LONGER
SLAVE	GRAVY	LAYER	EARLY	SPEAR	MARCH	STUNG	SPECK	PRIVY	EXTENT	DANGER
SUAVE	DROVE	LATER	FAIRY	SMEAR	PEACH	FLUNG	THICK	PRICK	INVENT	GAUGED
SCALE	BRAVO	LINER	DAIRY	STEAM	PUNCH	CLANG	BRICK	PRIOR	ORIENT	DODGER
SHARP	BROKE	SEWER	PERKY	CHEER	RANCH	ALONG	TRICK	POISE	CLIENT	NUGGET

Table A2 continues

__TTER	_A_ING	_INDER	__CKE_	__LLE_	_AR_ER	__BLE	RE__E	STR__	_EA_ER
BUTTER	RAVING	FINDER	POCKET	PALLET	BARKER	RUMBLE	RECITE	STRIKE	LEADER
MUTTER	PAVING	TINDER	PICKET	BALLET	BARBER	RUBBLE	REFUTE	STRIVE	READER
BITTER	SAVING	CINDER	WICKET	WALLED	BARREN	RAMBLE	REVISE	STRIDE	HEALER
PUTTER	HAVING	BINDER	ROCKET	CALLER	MARKER	HUMBLE	REVIVE	STRIFE	HEARER
BATTER	RACING	WONDER	SOCKET	POLLED	WARMER	MUMBLE	REFUSE	STRIPE	HEATER
LITTER	RATING	WANDER	TICKET	PULLED	LARDER	JUMBLE	RESIDE	STRAFE	BEAKER
MATTER	WAKING	FENDER	PACKET	FULLER	GARNER	FUMBLE	RELIVE	STROVE	BEAVER
SITTER	TAKING	GENDER	PICKED	WALLET	CARVER	RABBLE	REPUTE	STRODE	BEARER
BETTER	MAKING	WINNER	RACKET	FILLER	FARMER	TUMBLE	REFINE	STROKE	WEAVER
PATTER	SAYING	TENDER	WICKED	BILLED	GARTER	BUBBLE	RECIPE	STRING	HEADED
GUTTER	PALING	GANDER	LOCKED	VALLEY	WARDEN	GAMBLE	REGIME	STRONG	LEADEN
HITTER	FACING	WINTER	WICKER	FALLEN	BARRED	GABBLE	RETIRE	STRICT	DEALER
CUTTER	PAGING	GINGER	ROCKER	PULLEY	CAREER	NIBBLE	RESUME	STRAND	HEAVEN
LATTER	GAMING	SINGER	JACKET	BULLET	GARDEN	HOBBLE	REFUGE	STRUNG	WEAKEN
TITTER	GAPING	RENDER	LOCKER	ROLLER	WARREN	VIABLE	REMOTE	STRAIT	BEATEN
LETTER	LAYING	FINGER	TICKER	DULLER	HARDEN	LIABLE	REBUKE	STREAK	LEAFED
SETTER	EATING	YONDER	SICKER	FELLER	HARPER	ENABLE	RELATE	STREET	SEALED
POTTER	TAXING	SUNDER	COCKED	BILLET	BARBED	UNABLE	REDUCE	STRAIN	PEAKED
BUSTER	WANING	SINNER	BACKED	ROLLED	CARMEN	PEBBLE	REMOVE	STRATA	SEAMEN
MISTER	FADING	PONDER	HACKED	GALLEY	BARREL	MARBLE	REPOSE	STREAM	LOADER
MUSTER	DARING	LINGER	BUCKET	MILLER	BARLEY	USABLE	REMAKE	STROLL	WEASEL
MASTER	ROVING	DINNER	HOCKEY	POLLEN	MARKED	BAUBLE	REVERE	STRESS	GRADER
SISTER	RAVINE	WILDER	JOCKEY	VOLLEY	MARKET	STABLE	RECEDE	STRUCK	TRADER
PESTER	MOVING	WINDED	NICKEL	WELLED	BURDEN	EDIBLE	RESCUE	SHRINE	LOADED

_EA_ING	_AR_ING	I__ING	_O__ING	__A_TER	CON_E_T	___TURE	_A__AGE	S__PPER
LEADING	BARKING	WILLING	COOLING	CLATTER	CONTENT	VENTURE	VANTAGE	SHIPPER
HEADING	WARNING	WINDING	BOOKING	FLATTER	CONCERT	GESTURE	BANDAGE	SLIPPER
HEARING	WARPING	FINDING	FOOLING	PLATTER	CONSENT	FEATURE	WASTAGE	SKIPPER
LEAVING	WARRING	FILLING	TOOLING	SHATTER	CONTEST	TEXTURE	PASSAGE	SHOPPER
HEATING	MARKING	WINKING	FOOTING	CHATTER	CONVENT	LECTURE	BARRAGE	CLIPPER
BEARING	BARRING	WINNING	ROLLING	SPATTER	CONNECT	RUPTURE	MASSAGE	SNAPPER
LEANING	WALKING	MILLING	ROOFING	PLASTER	CONVERT	PASTURE	YARDAGE	STOPPER
HEALING	PARTING	SINKING	BOATING	SCATTER	CONTEXT	CULTURE	GARBAGE	SLIPPED
LEAPING	EARNING	SINGING	BOOMING	FLUTTER	CONCEPT	CAPTURE	SAUSAGE	FLAPPER
READING	BACKING	KILLING	LOOKING	PLANTER	CONTACT	POSTURE	SALVAGE	CHOPPER
HEAVING	DARLING	LISTING	BOLTING	CHARTER	CONCERN	RAPTURE	PACKAGE	SHARPER
SEARING	HARPING	WISHING	BOILING	CHANTER	CONTEND	MIXTURE	HAULAGE	WRAPPER
LEASING	CARVING	BINDING	ROARING	CHAPTER	CONVICT	VULTURE	BAGGAGE	SLEEPER
WEARING	FARMING	MISSING	COATING	SHUTTER	CONDUCT	FIXTURE	VINTAGE	WHIMPER
GEARING	BANKING	MINCING	MOORING	GLITTER	CONSIST	PICTURE	CABBAGE	TRAPPER
BEATING	BASKING	FISHING	ROCKING	CLUSTER	CONSORT	NURTURE	VOLTAGE	WHISPER
SEATING	VARYING	SIBLING	POLLING	STARTER	CONSULT	TORTURE	PORTAGE	STOPPED
MEANING	PACKING	KISSING	ROUSING	SPUTTER	CONCEDE	STATURE	HOSTAGE	STEPPED
DEALING	WANTING	FITTING	HOWLING	BLUSTER	CONCEAL	CENSURE	BONDAGE	WHIPPET
LENDING	WAITING	SITTING	ROMPING	FLATTEN	CONFESS	MEASURE	RAMPAGE	CLAPPED
BENDING	SACKING	TILTING	POTTING	QUARTER	CORRECT	LEISURE	COTTAGE	SHIMMER
HERRING	WAILING	HISSING	SOAKING	ADAPTER	COLLECT	RESTORE	LINKAGE	SLICKER
MENDING	WASHING	DIALING	HOUSING	BLISTER	CONDEMN	SEIZURE	LINEAGE	SKINNER
LOADING	WORKING	LIFTING	ROUTINE	FLOATER	COMMENT	CENTURY	VILLAGE	SLENDER

Table A2 continues

S_E__ING	___N_ING	___ATION	___PPING	____LESS	OVER____
SWEEPING	POINTING	ROTATION	TRIPPING	HANDLESS	OVERHAND
SLEEPING	POUNDING	DONATION	TRAPPING	HEEDLESS	OVERLAND
SHEETING	SOUNDING	NOTATION	SHIPPING	TIRELESS	OVERLOAD
STEERING	PRINTING	VOCATION	WRAPPING	TIMELESS	OVERTAKE
SNEEZING	MOUNTING	RELATION	WHIPPING	LIFELESS	OVERLAID
SWEATING	COUNTING	DILATION	SHOPPING	BASELESS	OVERHEAD
SWEARING	BOUNDING	LOCATION	DRIPPING	HOPELESS	OVERCOME
SPEAKING	GRUNTING	VACATION	CHIPPING	HEADLESS	OVERDONE
SHEARING	PAINTING	CITATION	DROPPING	CARELESS	OVERPAID
SNEAKING	ROUNDING	NEGATION	WHOPPING	NAMELESS	OVERHANG
SWELLING	STUNNING	DURATION	CHOPPING	WINDLESS	OVERRODE
SPELLING	PLUNGING	TAXATION	SNAPPING	TONELESS	OVERCOAT
SLEDDING	GRINDING	GYRATION	SLAPPING	MINDLESS	OVERTIME
BREEDING	STINGING	CREATION	STOPPING	HARMLESS	OVERHAUL
SPENDING	STINKING	AVIATION	THUMPING	NEEDLESS	OVERHEAT
SMELLING	BOUNCING	VOLITION	GRASPING	HAIRLESS	OVERCAST
GREETING	TAUNTING	EQUATION	GROUPING	LOVELESS	OVERRIDE
STERLING	SWINGING	SOLUTION	TRAILING	LISTLESS	OVERLOOK
BLEEDING	SPINNING	DILUTION	TRAINING	PEERLESS	OVERTOOK
CREEPING	PLANTING	PETITION	TRIMMING	HELPLESS	OVERFEED
FREEZING	GRINNING	POSITION	TRIFLING	VIEWLESS	OVERFALL
FLEETING	STANDING	DEVOTION	THINKING	PITILESS	OVERSIZE
GREENING	DRINKING	REACTION	THRIVING	PAINLESS	OVERTURE
BREAKING	PLANKING	SANCTION	THROWING	RESTLESS	OVERSEAS

Table

Table 1. Averages over participant's new and old Z values for Experiment 3, with corresponding old response probabilities in brackets.

	Dissimilar				Similar		
Category Length	0	4	8	12	4	8	12
New	-1.151	-1.203	-1.308	-1.395	-0.676	-0.436	-0.414
	(.125)	(.114)	(.095)	(.082)	(.250)	(.331)	(.339)
Old	0.802	0.784	0.753	0.746	0.895	0.862	0.957
	(.789)	(.783)	(.774)	(.772)	(.815)	(.806)	(.831)

Figure Captions

Figure 1. Fits of the dual-process model to data from Glanzer et al.'s (1999a) Experiment 1 measured from Yonelinas's (1999) Figure 3.

Figure 2. Experiment 1 (semantic similarity): Average ZROC functions for the unequal-variance normal model (solid lines) fit separately to dissimilar and similar words, and 95% confidence intervals on the average deviations of the observed Z scores from the Z scores predicted by the model, with means joined by dotted lines. *D'* is intercept/slope, and all estimates are the average of individual participant's estimates.

Figure 3. Experiment 2 (orthographic similarity): Average ZROC functions for the unequal-variance normal model (solid lines) fit separately to dissimilar and similar words, and 95% confidence intervals on the average deviations of the observed Z scores from the Z scores predicted by the model, with means joined by dotted lines. *D'* is intercept/slope, and all estimates are the average of individual participant's estimates.

Figure 4. Experiment 3: Average ZROC functions for the unequal-variance normal model (solid lines) fit separately to dissimilar and similar words at each category length, and 95% confidence intervals on the average deviations of the observed Z scores from the Z scores predicted by the model, with means joined by dotted lines. *D'* is intercept/slope, and all estimates are the average of the individual participant's estimates.

Figure 5. Experiment 4: Average ZROC functions for the unequal-variance normal model (solid lines) fit simultaneously to dissimilar and similar words and separately to the first and last two blocks, and 95% confidence intervals on the average deviations of the observed Z scores from the Z scores predicted by the model, with means joined

by dotted lines for dissimilar words and dash-dot lines for similar words. Parameters for the equations are the average of individual parameters.

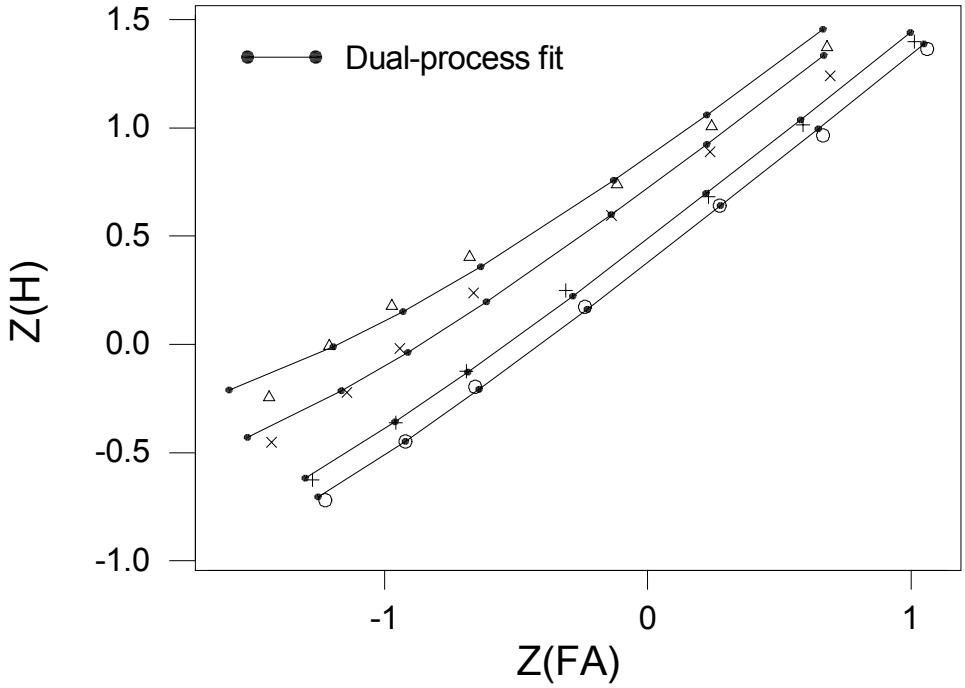
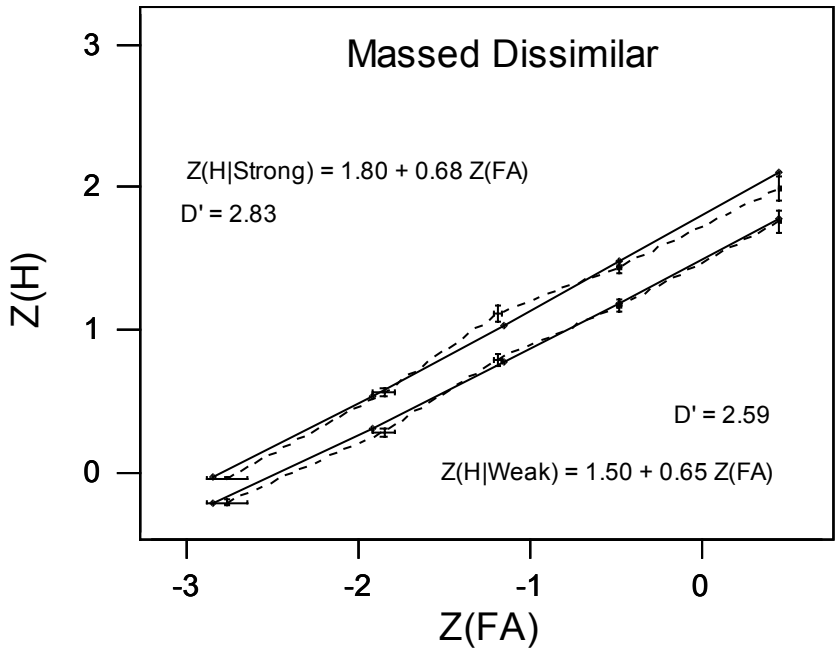
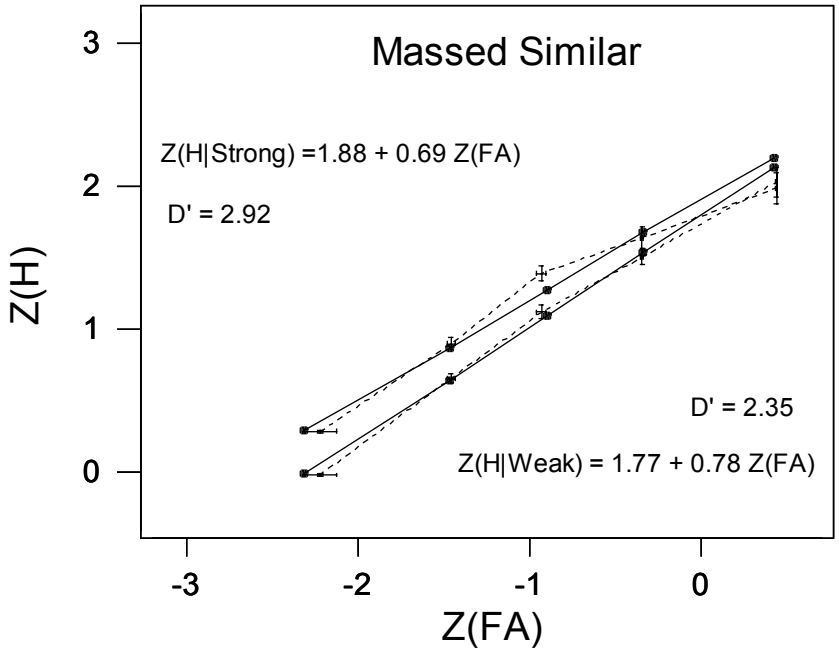


Figure 1

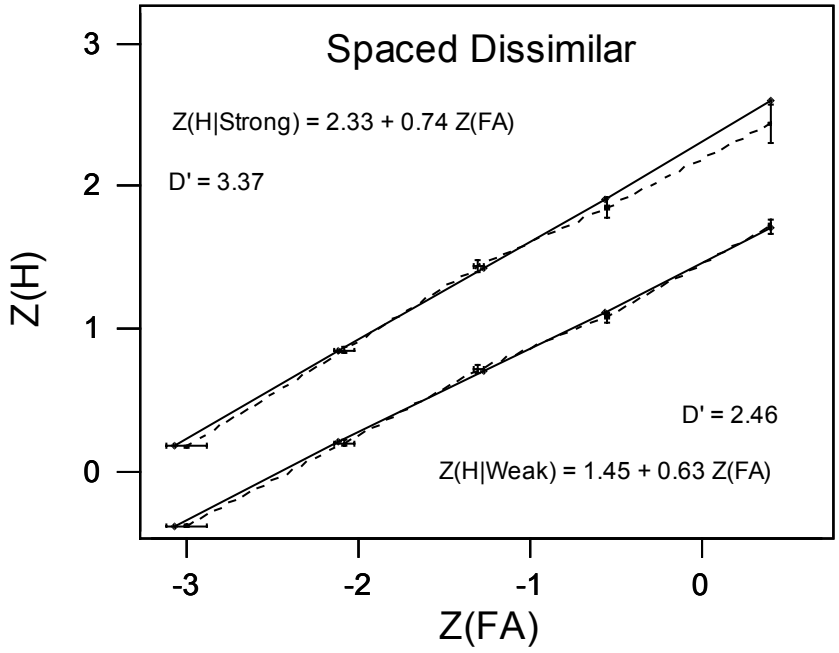

Massed Dissimilar
Z(H|Strong) = 1.80 + 0.68 Z(FA)
D' = 2.83
D' = 2.59
Z(H|Weak) = 1.50 + 0.65 Z(FA)
Z(H)
Z(FA)
3
2
1
0
-3
-2
-1
0

(a)

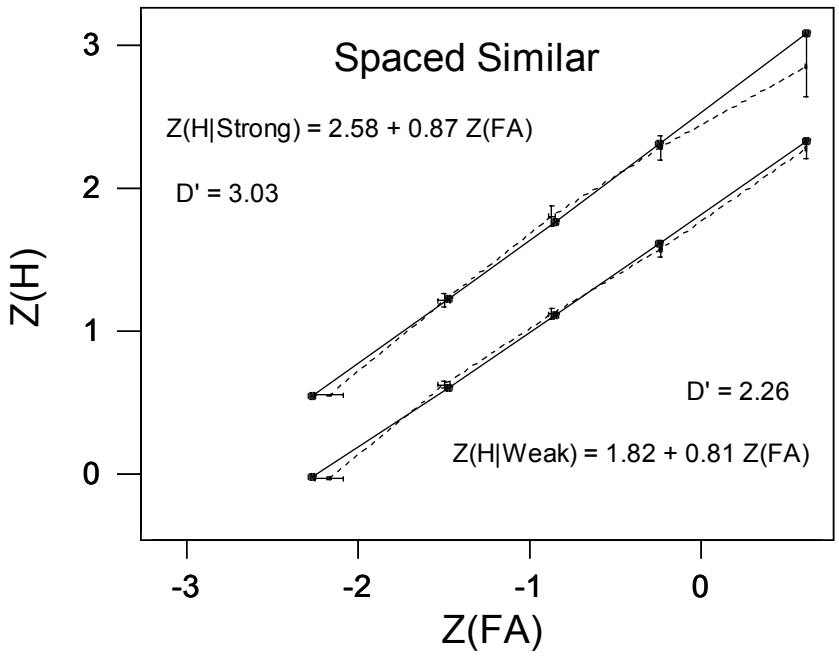

Massed Similar
Z(H|Strong) =1.88 + 0.69 Z(FA)
D' = 2.92
D' = 2.35
Z(H|Weak) = 1.77 + 0.78 Z(FA)
Z(H)
Z(FA)
3
2
1
0
-3
-2
-1
0

(b)

Figure 2

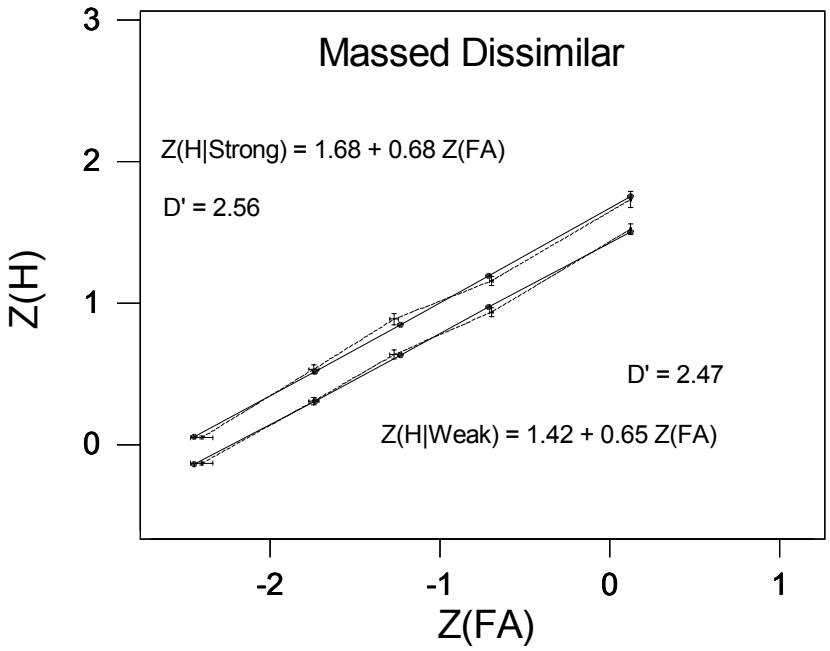


(c)

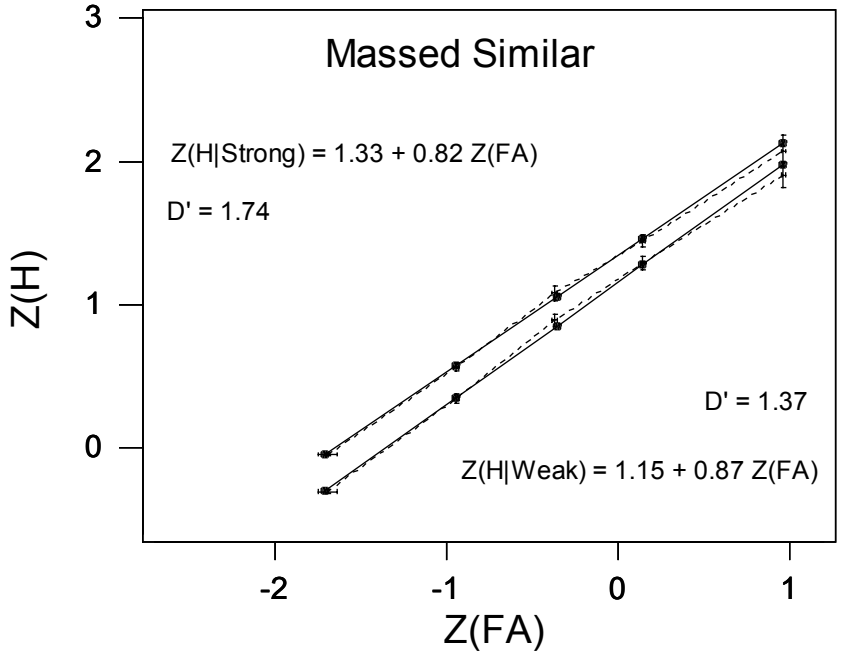


(d)

Figure 2 (continued)

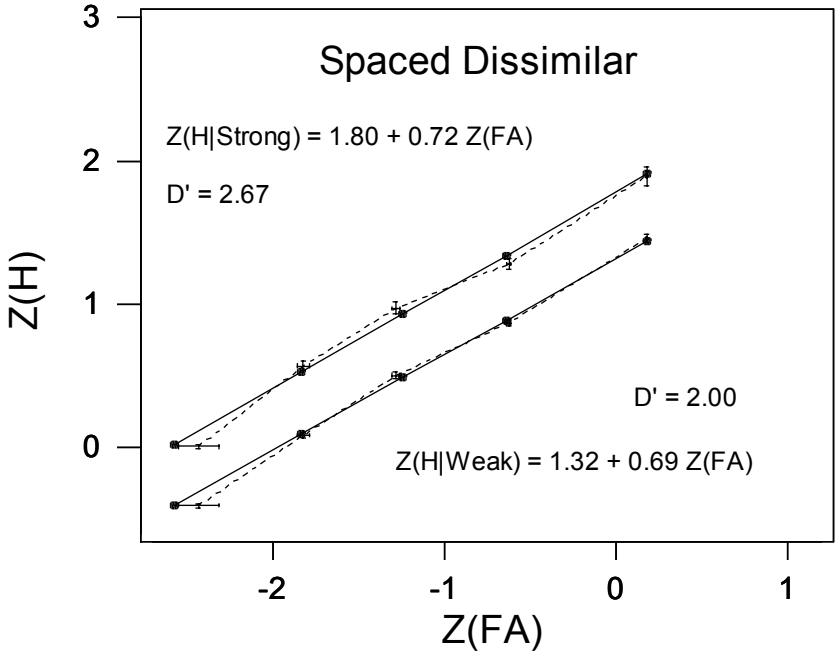

Massed Dissimilar
Z(H|Strong) = 1.68 + 0.68 Z(FA)
D' = 2.56
D' = 2.47
Z(H|Weak) = 1.42 + 0.65 Z(FA)
Z(H)
Z(FA)
3
2
1
0
-2
-1
0
1

(a)

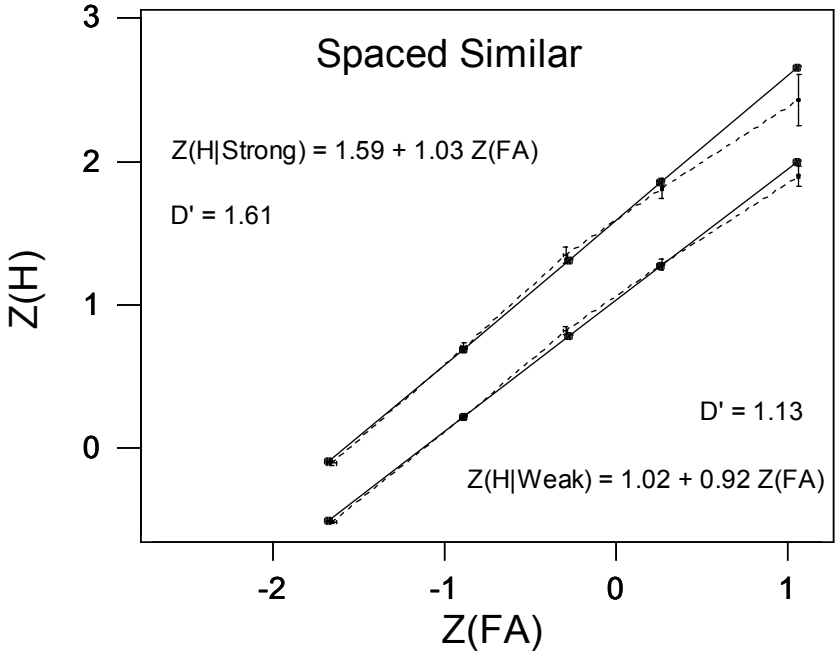

Massed Similar
Z(H|Strong) = 1.33 + 0.82 Z(FA)
D' = 1.74
D' = 1.37
Z(H|Weak) = 1.15 + 0.87 Z(FA)
Z(H)
Z(FA)
3
2
1
0
-2
-1
0
1

(b)

Figure 3

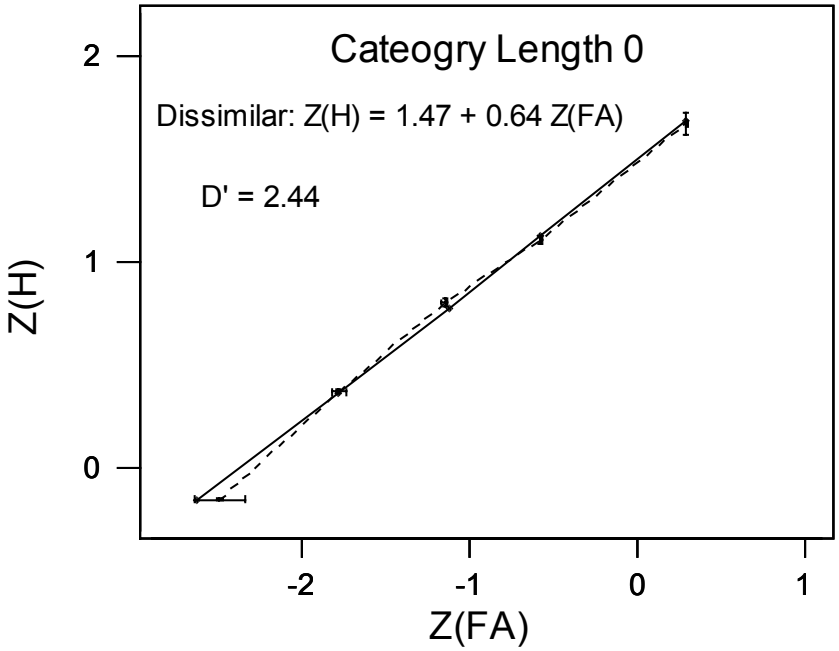


(c)

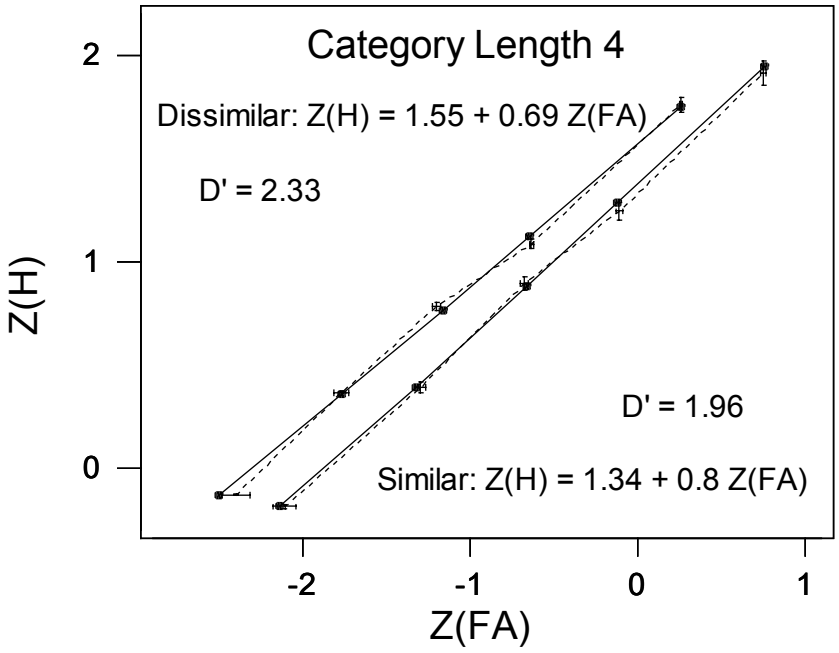


(d)

Figure 3 (continued)

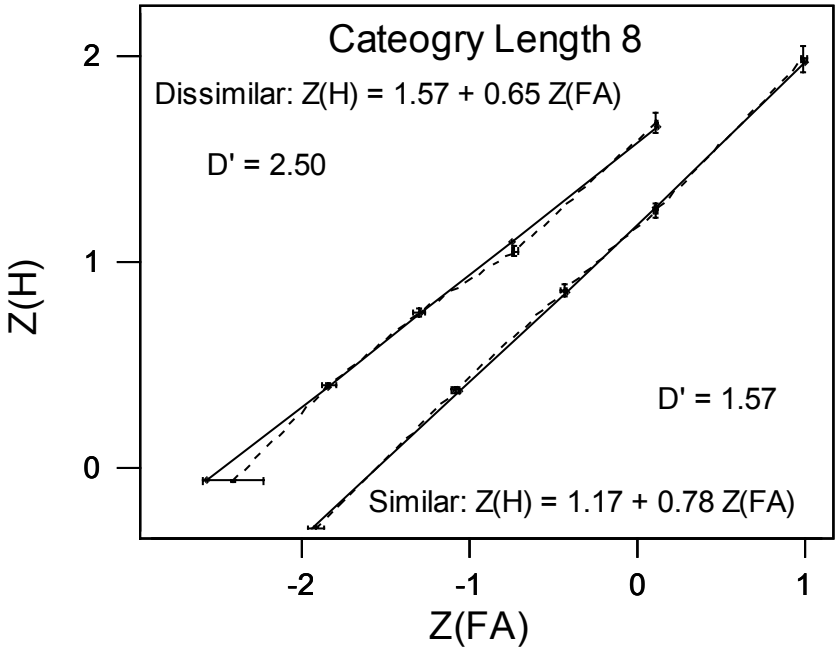

Cateogry Length 0
Dissimilar: Z(H) = 1.47 + 0.64 Z(FA)
D' = 2.44
Z(H)
Z(FA)
2
1
0
-2
-1
0
1

(a)

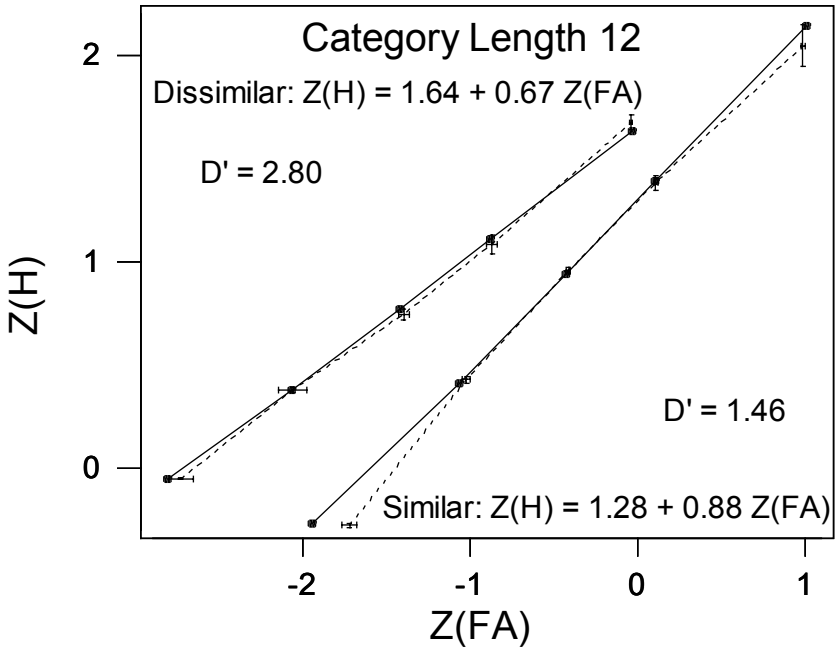

Category Length 4
Dissimilar: Z(H) = 1.55 + 0.69 Z(FA)
D' = 2.33
D' = 1.96
Similar: Z(H) = 1.34 + 0.8 Z(FA)
Z(H)
Z(FA)
2
1
0
-2
-1
0
1

(b)

Figure 4

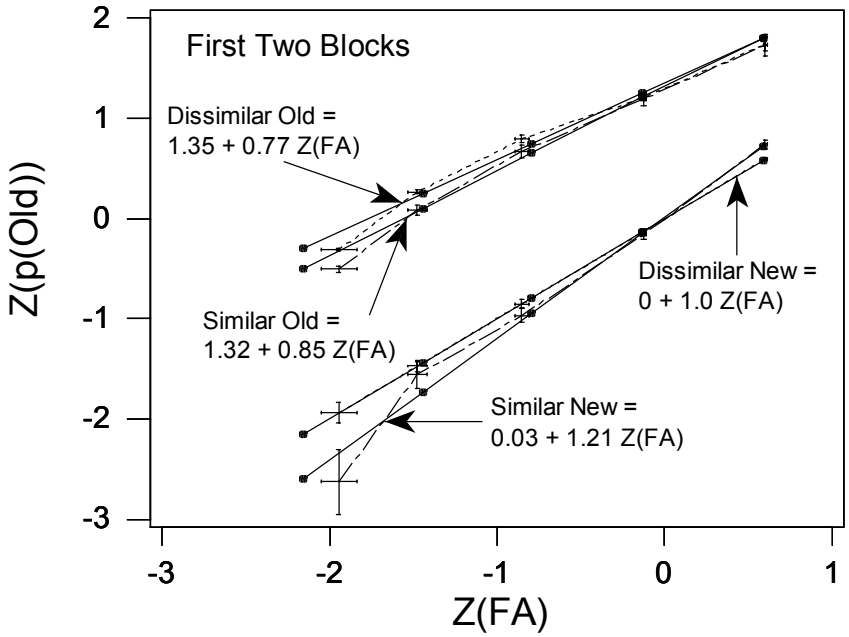


(c)

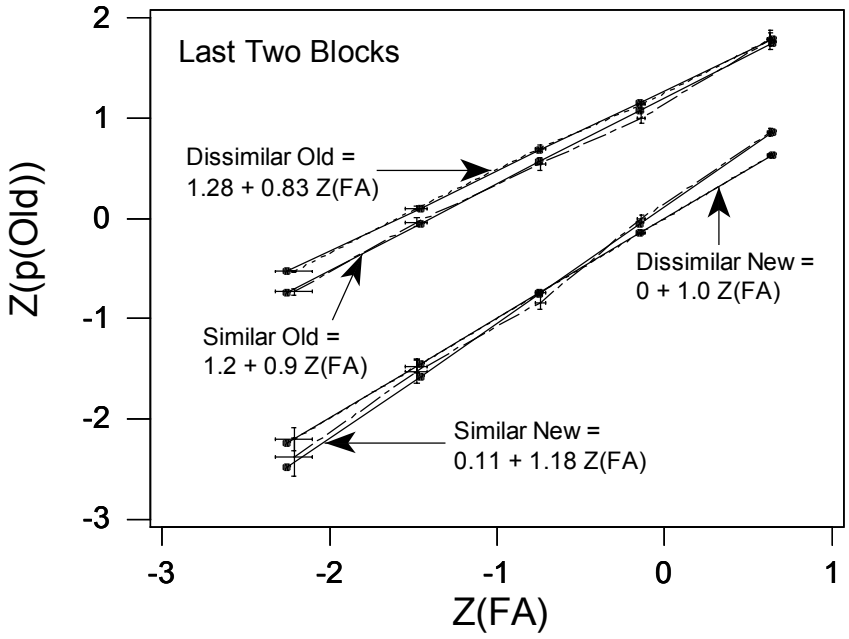


(d)

Figure 4 (continued)


First Two Blocks
Dissimilar Old =
1.35 + 0.77 Z(FA)
Similar Old =
1.32 + 0.85 Z(FA)
Dissimilar New =
0 + 1.0 Z(FA)
Similar New =
0.03 + 1.21 Z(FA)
Z(p(Old))
Z(FA)
2
1
0
-1
-2
-3

(a)


Last Two Blocks
Dissimilar Old =
1.28 + 0.83 Z(FA)
Similar Old =
1.2 + 0.9 Z(FA)
Dissimilar New =
0 + 1.0 Z(FA)
Similar New =
0.11 + 1.18 Z(FA)
Z(p(Old))
Z(FA)
2
1
0
-1
-2
-3

(b)

Figure 5