What is and what does free association measure?

Douglas Nelson
Cathy McEvoy
University of Southern Florida

Simon Dennis
University of Queensland

Memory tasks and their theoretical implementations model everyday memory tasks. Research on free recall, cued recall, and recognition as well as work with implicit memory procedures capture some if not the majority of the characteristics of the memory requirements needed in interactions with the environment. In this context, it is interesting that a memory task that has been used in research for more than one hundred years (Galton, 1880) has not attracted much theoretical attention. We refer to free association, a task that requires participants to produce the first word to come to mind that is related in a specified way to a presented cue (e.g., meaning, rhyme, makes a word). This task is used in everyday activities as a means for "collecting thoughts." For example, in using the yellow pages free associating to a needed product or service can be helpful in determining an effective search heading. The same advantage is apparent when attempting to find information on the internet, or just the right word in writing and speech. Similarly, we are often called upon to provide information in response to broad and unanticipated questions such as "Why are you a democrat?" In such cases, recovering the needed information can be more akin to free association to the question than to directed recall.

Free association is a utilitarian production task that has been directed to a variety of different types of information, such as categories, words, rhymes, word stems and fragments

(e.g., Battig & Montague, 1969; Graf & Williams, 1987; Nelson & Brooks, 1974; Nelson, Schrieber, & McEvoy, 1992; Palermo & Jenkins, 1964; Postman & Keppel, 1970). What we find interesting is the complete absence of attempts to model the nature of free association as a memory task. This deficiency is puzzling given the current interest in implicit memory which often depends on contrasting free association and cued recall instructions in order to study implicit and explicit memory processes (e.g., Graf & Mandler, 1984; Nelson, Schreiber, & Holley, 1992; Roediger, 1990; Weldon & Coyote, 1996). This deficiency is also surprising given success with free association in predicting cued recall (Bahrick, 1970; Nelson, McKinney, Gee, & Janczura, 1998), alcohol and marijuana use (e.g., Stacy, 1997), and in showing memory biases in depression (e.g., Watkins, Vache, Verney, & Mathews, 1996) and in pain patients (Pincus, Pearce, McClelland, Farley, & Vogel, 1994).

We suspect that there are three main reasons for this failure. First, the history of free association is linked to scientific misadventures. This method provided the database for the British empiricists who relied on introspection to discover the laws of association (e.g., Anderson & Bower, 1974; Deese, 1965) and, in addition, it has had a checkered history as a diagnostic tool for clinicians (Deese, 1965). Second, researchers have no control over the conditions surrounding original acquisition and subsequent use in language. Controlling the conditions of encoding has served as a benchmark for laboratory research since Ebbinghaus. Third, and we think importantly, free association in recent years has been treated as a yardstick in the traditional measure-then-control-or-manipulate approach to research. Free association is used to select cues for other words in order to control or to manipulate such variables as strength of prior association, set size, and so on (e.g., Nelson et. al., 1998) just as printed

frequency norms (e.g., Kucera & Francis, 1967) have been used to accomplish similar ends (e.g., Gillund & Shiffrin, 1984). Using free association as a yardstick for assessing performance in other situations seems to have overshadowed the possibility of investigating the yardstick itself. Unlike printed frequency norms in which words are counted across samples of text, words are produced in free association and we know little about the representations and processes involved.

The present paper presents and evaluates a model of free association that grew out of attempts to understand findings initially generated to examine free association as an index of strength. In discrete tasks only a single response is produced for each normed word or cue with strength of association traditionally measured by counting the number of people who produce each response. This count is then divided by the sample size to determine response probability. These probabilities form a hierarchy of strengths from strongest to weakest, but despite the usefulness of the strength index in predicting performance, it comes without estimates of error. Nevertheless, if the index is reliable, this shortcoming is not a problem for most purposes, especially when strength is being controlled or being manipulated categorically (e.g., strong versus weak). However, although there are many sets of norms (e.g., Brown, 1976; Cramer, 1968; Nelson, McEvoy & Schreiber, 1998), finding published information about reliability proved difficult. Although early work on free association indicated that reliabilities were in the high nineties, these studies were based on continuous association using nonsense syllables as cues almost exclusively (Cieutat, 1963; Gaeth & Allen, 1966; Johnson, 1968; Noble, 1952; also see Simpson & Voss, 1967 who used words). Continuous association is problematic because it is subject to both response chaining and retrieval inhibition (McEvoy &

Nelson, 1982). In addition, reliability was generally calculated for the number of associations per word, what is now called set size, not strength. We found only two references that specifically addressed the reliability of the strength index using discrete free association (Jenkins & Palermo, 1965; Nelson & Schreiber, 1992).

The issue of reliability should be of equal concern to those using the strength index to control or to predict performance. It should also be of interest to those using other procedures to assess connections between words and claim that such procedures will predict free association performance (e.g., Landauer & Dumais, 1997). Reliability serves as a limit on what can be achieved in either case. Just as importantly, in the absence of a model of free association, the issue of validity also deserves deeper consideration. Free association ostensibly provides a valid measure of pre-existing strength between a test cue and a related word, and strength simply means that, given a particular cue, one response is more likely than another. The term strength in this context is theoretically neutral and little consideration has been given to the nature of the representations that underlie the construct. Strength of a response could be implemented as the number of its instances in memory, with stronger associations reflecting larger numbers of instances (Hintzman, 1976; Jacoby & Brooks, 1984; Landauer, 1975), or it could be implemented as a single abstract representation that increases in potency with experience (e.g., Noble, 1952; Rosen & Russell, 1957). The idea that free association provides an index of what was called "habit strength" in the 1950s and 60s probably best captures what some investigators who use norms assume, if only implicitly, and this assumption could be said to carry the supposition that free association indexes absolute strength. For example, Rosen and Russell implied absolute strength when they suggested that *table* was 2.5 times stronger

than *sit* as a response to the cue CHAIR. In the norms, the probabilities of *table* and *sit* were, respectively, .50 and .20 and .50 is 2.5 times as large as .20. This supposition seems strained at best because the very nature of the procedure is relative. The more people who produce the primary response (the strongest or most popular response), the fewer people there are who can produce weaker responses.

These considerations suggest that free association can be a successful predictor and may be an interesting memory task in its own right, but also that there are ambiguities concerning the reliability and validity of its main index. To explore these issues, we adopted a procedure used by Rosen and Russell (1957). The standard discrete association task was modified by asking participants to provide two associates for each to-be-normed word instead of a single word. They were asked to produce the first word to come to mind on each occasion but not to repeat the same word. This procedure had four advantages. First, because only two responses were requested problems with retrieval inhibition are minimized relative to continuous production. In addition, the potential for response chaining can be evaluated by determining whether the first response is associatively related to the second. Second, by correlating the probabilities of words produced as first-responses with probabilities collected in earlier norms, the reliability of the strength index can be assessed. In addition, reliability can also be assessed within the same sample of participants by correlating the probabilities for a given response as first and as second responses. For example, the word BANK produced six responses in common as first and as second associates, and these items and their respective first-second probabilities were: *money* (.75, .54), *robbery* (.07, .09), *account* (.05, .02), *teller* (.03, .10), *deposit* (.03, .06), and *book* (.01, .01). Third, collecting the dual-response norms allowed us

to examine the reliability of free association as a means for estimating set size which refers to the number of responses or associates linked to a given word. Studied words with fewer associates tend to be easier to reproduce in cued recall, and such words are more effective as test cues (e.g.,. Nelson, Schreiber, & McEvoy, 1992). An earlier study indicated that free association provided a reliable index of set size (Nelson & Schreiber, 1992), but this assessment was based on re-norming words with independent samples of participants. The dual-response procedure offers the opportunity to determine whether set size estimated from the first response is similar to set size estimated from a second response. Reliability will be higher when the same responses are given on both occasions and it will be lower when new responses are introduced.

Fourth, collecting the dual-response norms allowed us to evaluate two different conceptualizations of the strength index as assessed by free association. We asked participants not to repeat their first response in order to determine how the probabilities change when the first response was unavailable. If response probability indexes absolute strength, then probability is strength and the ratio of the probability of the first response to the probability of the second response should approach 1.0 at each associative rank (primary associate, secondary associate, and so on). To continue the BANK example, the probability of producing money, robbery, and so on, is predicted to be the same regardless of whether the item is produced as a first or as a second response. The second conceptualization, the relative strength hypothesis, makes a different assumption about the meaning of strength in free association. According to the relative hypothesis, probability of response production is a manifestation of strength, not strength itself. The relative hypothesis assumes that reading a to-be-normed word

in conjunction with instructions about how to respond to it produces a parallel activation of associates in the specified domain. The strength of any given response in this domain is not represented by a single value but by a distribution of values and, for a given subject at a given moment, the strongest associates in this pool are selected and reported. Response selection is based on sampled strengths relative to other associates in the set. This hypothesis was used as the motivation for a simulation model described later. The important point for now is that the model used the first response probabilities to characterize the strength distributions of each response and then estimated the second response probabilities. Although the implications of the absolute and relative strength hypotheses were evaluated by different means, the second response probabilities were crucial to the evaluation of each idea.

Three different types of items were normed. One-third of the words had small sets of associates with strong primaries, another third had small sets of associates with weak primaries, and the final third had large sets of associates with weak primaries. These item-types were selected because their response distributions were grossly different and therefore more likely to challenge our earlier conclusions about the reliability of strength and set size measures (Nelson & Schreiber, 1992) as well as the proposed hypotheses about what the strength index measures.

Method

Materials. The to-be-normed items consisted of 120 words selected from single-response norms collected earlier for a pool consisting of 5,000+ words (Nelson, McEvoy, & Schreiber, 1998; Nelson & Schreiber, 1992). Each of the 120 words is presented in Appendix A. As determined by the single-response norms, 40 of these words had small associative sets

(5.13, SD = 1.26) with stronger primary associates (.81, SD = .04). Each word had 8 or fewer associates produced by two or more participants, and an average of 81% of the subjects produced the primary associate. Similarly, 40 of the 120 words had small associative sets (7.32, SD = 1.02) with weaker primary associates (.37, SD = .05), and 40 had large associative sets (21.33, SD = 2.12) with weaker primaries (.30, SD = .04). Items with large sets and strong primaries (greater than or equal to .75) were not included in this study because such items cannot be found in the normative sample.

Other than differences in set size and primary strength, the 120 words were fairly similar to each other and to the larger sample of 5,000+ items evaluated in the single-response norms. Eighty-nine percent of the words were nouns, 5% were verbs and 5.8% were adjectives. No attempt was made to control concreteness, printed word frequency (Kucera & Francis, 1967), or the non-response index (omissions) but these values tended to be close across item-types, respectively averaging, 5.34 (SD = 1.12), 40 words per million (SD = 79), and .02 (SD = .03). In comparison to the larger database, the words selected for this study tended to be somewhat more concrete (5.34 versus 4.75), less frequent (40 versus 77), and less likely to produce non-responses (.02 versus .03). In addition, approximately 39% (SD = .23) of the associates of these normed words produced that word as an associate when normed separately, and each word had an average of 1.43 (SD = .71) connections among its associates. These values were well within a standard deviation of the larger database of 5,000+ words (.39 versus .30 and 1.43 versus 1.62). In general, the words selected for the present study were concrete nouns occurring with moderate frequencies in text, and their connective organization approximated the general pool of words normed in the larger sample.

Participants. The 300 participants were undergraduates in introductory psychology who participated on a voluntary basis on the first day of class.

Procedure. Each participant was given one of two booklets containing 60 of the to-be-normed words, 20 for each word-type, so that a total of 150 responded to each item. This was the same number of participants on average that were involved in collecting the single-response norms and all norms were collected from introductory courses. Each booklet contained instructions on the first page and the stimulus words appeared on the following 3 pages. Both the order of the item-types and the order of the pages were unsystematically randomized for each subject. Related words (e.g., emerald and jewel) were assigned to different booklets.

For the dual-response norms, each cue was typed twice in a row with an underlined space appearing next to each word and its repetition. Participants were asked to read the words appearing on the left before writing the first word to come to mind in the space, and importantly, they were asked not to repeat their first response when responding to the same word for a second time. With this procedure, each item generated a first and a second response and the strengths of given responses could be determined for each associative rank. Normally, in single-response free association norms, the responses are arranged in an ordered array determined by the probability that a particular response is produced, e.g., the primary associate, secondary associate, and so on. The same procedure was followed for the dual-response norms except that the presence of two responses was used to create two arrays, one for the first and one for the second response. With this procedure, the probability of producing *car* to the cue ACCIDENT could be determined as a first response and as a second response.

Following tradition, the probability of any given response was determined by counting the number of subjects who produced it divided by the sample size. Because repetitions were not allowed by the instructions, the sample size for the second response included only participants who did not produce that response as a first response. Finally, set size was determined by counting the number of different responses made by two or more people, and this count was done separately for both the first and second response sets.

Results

Free association as an index of strength and set size

First response norms and strength. The reliability coefficients were determined by calculating the correlations between the probability of a given response in the single-response norms with the probability of the same word in the dual-response norms. In the dual-response norms, only first response data were used. Reliability was high for the set as a whole, r = .89, and this value corroborated the results (r = .89) of an earlier study of 155 items (Nelson & Schreiber, 1992). These findings suggest that free association strengths obtained in the discrete task and indexed by response probability are stable over different groups of participants (the reliabilities for individual items are shown in Appendix A). Interestingly, 19 words in the set could be classified as homographs and the reliability of these items was close to the reliability of the set as a whole, r = .90.

The reliability for the three item-types were, respectively, .99 (SD = .01), .87 (SD = .14), and .82 (SD = .22) for words having small set-strong primaries, small set-weak primaries, and large set-weak primaries. Although the lowest coefficient was reasonably high, reliability

clearly differed as a function of item-type with small-strong items showing the highest reliability and large-weak items showing the lowest values on average. An items nested in item-type analysis of variance indicated that these differences were significant, $F(2, 117) = 14.57$, $MSe = .023$, and a Fisher's LSD of .07 indicated that reliability was significantly greater for items having small sets and strong primaries than for the other two item-types which did not differ. A re-analysis by item-type of our earlier reliability study (Nelson & Schreiber, 1992) confirmed the present findings. Although the number of items differed widely, reliability was highest for small-strong items (.99, n = 13), next for small-weak items (.84, n = 7) and lowest for large-weak items (.81, n = 40). The majority of the items normed at that time had medium sized sets, and their reliability also tended to be high (.89). Hence, words with exceptionally strong primary associates were more likely to reproduce their associates at the same levels of strength than other types of words but, despite this difference, the strength index tended to show high levels of reliability for the majority of the items of all types.

Dual-response norms and strength. Reliability was also determined by computing the correlations between the probabilities of a given item as the first and as the second response in the dual-response norms. The probability of the second response was corrected for opportunity because participants were asked not to repeat their first response. To clarify this correction, consider the cue EMPLOYER and one of its responses, *boss*: 45/149 participants produced *boss* as their first response for a probability of .30, and 28/104 participants (149 - 45 = 104) produced *boss* as their second response for a conditional probability of .27. The reliability coefficients were then computed on these probabilities which remained fairly high overall at .83. Reliability was highest for small-strong items (.90, SD = .12), next highest for large-weak items

(.82, SD = .16), and lowest for small-weak items (.76. SD .24). Reliability was significantly greater once again for small-strong than for the other items, $F(2, 117) = 6.22$, $MSe = .028$, $LSD = .08$.

Although the reliabilities for the first and second responses tended to be high, they were somewhat lower than for "first" responses when set size was small. The reliability for small-strong items declined from .99 for first responses to .90 for first-and-second responses and, similarly, for small-weak items it declined from .87 to .76. This change suggests that, at least for words with smaller sets, the probability of the second response tended to differ from the probability of the first response. To examine this change in more detail, the mean response probabilities for the six most frequent associates were calculated for each item-type for the first and second responses. The rank ordering of the six associates was determined by the responses produced first.

The results of these calculations shown in Table 1 were subjected to an items nested in item-type mixed-model analysis of variance with order (first versus second) and rank serving as within-subject variables. As expected, item-type had a reliable effect, $F(2, 117) = 77.56$, $MSe = .004$, because the associates of small-strong items (.14) and small-weak items (.14) were stronger than the associates of large-weak items (.09). This effect was a predictable byproduct of the free association procedure. Also, as expected, rank was significant, $F(5, 585) = 612.61$, $MSe = .007$, as was the interaction between item-type and rank, $F(10, 585) = 67.23$. This interaction can be mainly attributed to small-strong items, which by definition, have exceptionally strong primary associates.

Insert Table 1 about here

What was more interesting in this analysis, the strength of any given response tended to be greater when it was produced as the first response (.13) than when it was produced as the second response (.11), $F(1, 117) = 87.41$, $MSe = .002$. Although the overall effect was numerically small, large differences were apparent for some of the item-types. The effects of response order interacted with item-type, $F(2, 117) = 5.04$, rank, $F(5, 585) = 137.46$, $MSe = .003$, and the three-way interaction among these variables was also reliable, $F(10, 585) = 24.08$. The interaction with rank is shown in the last row of Table 1 and, as can be seen, response order effects were most apparent for the primary associate (Rank 1). The probability of this associate was substantially greater when it was produced as the first response than when it was produced as the second response, and as suggested by the three-way interaction (LSD = .02), this difference was most apparent for the small-strong items. Beyond Rank 1, the differences between first and second responses were much smaller.

Set size. Table 2 presents mean set sizes for each item-type for both the earlier single-response norms and the present dual-response norms. Values for the latter norms are shown for the first response, the second response, and for pooled responses based on words common to the two responses. As shown in the first two columns of Table 2, mean set size for each item-type tended to be stable. For example, for small-strong items, set size averaged 5.12 words in the single-response norms and 6.43 for the first response in the dual-response norms. Set size reliability of the associates produced by two or more participants in the two sets of norms averaged $r = .85$ for the 120 items. This relationship closely matched the results found in

an earlier study in which the correlation was r = .84 (Nelson & Schreiber, 1992). Interestingly, set size reliabilities derived from the single-response norms and both the second response (.46) and the pooled response (.66) indices were substantially lower. The correlation between the first and second responses of the dual-response norms was even lower, r = .39. The reliability of the set size index was clearly higher for the first than for the second response.

Insert Table 2 about here

Examination of the means shown in Table 2 suggested that the reason for the instability in the second response index could be traced to the increasing numbers of new words. The estimates of set size increased dramatically, particularly for small-strong items. For example, for small set-strong primary items, mean set size increased from 6.43 to 19.08. Given that the pooled response data closely approximated the second response data, the increase in estimated set size for the second response appeared to be the result of the production of new words not produced as first responses. An items nested in item-type mixed model analysis of variance that included response order as a variable indicated that these effects were significant. Set size varied with item-type, $F(2, 117) = 122.38$, $MSe = 10.72$ as expected. More importantly, set size was significantly larger for second than for first responses, $F(1, 117) = 424.37$, $MSe = 8.46$, and this effect varied with item-type, $F(2, 117) = 43.67$, $LSD = 1.27$. Set size increased significantly for each item-type, and whereas this increase was approximately equal for small-weak and large-weak items, it was especially large for small-strong items.

These results could mean that the single response procedure provides a good index of a word's strongest associates but not its weakest associates which tend to be produced only when additional response opportunities are provided. Alternatively, this procedure may provide a good estimate of set size for both the strongest and weakest responses, but asking for a second response distorts the estimate because such responses are not independent of the first. The second response could have been produced by the first response rather than by the normative cue (or both). Such response-to-response production is known as "chaining," and additional analyses were conducted to determine which alternative was more likely.

First, we examined the strength and number of new words that appeared in the dual-response norms that did not appear in the single-response norms. Table 3 presents the results of this analysis. The mean strengths shown in the first column indicate that new items were very weak, averaging .014 (produced by 2-3 participants). Although the mean strengths appeared to be uniform across item-types, an analysis of variance indicated that item-type was a significant source, $F(1, 117) = 8.05$, $MSe = .00001$, $LSD = .002$. Although the effect was meager, small-strong items produced significantly stronger new items than either small-weak or large-weak items which did not differ. The same patterns were found for the mean numbers of new words, but the differences were more apparent. As shown in the second column, more new words were produced in the small-strong condition than in either of the other item-type conditions, $F(1, 117) = 12.94$, $MSe = 18.27$, $LSD = 1.89$. When given a second response opportunity, new words tend to be very weak and small-strong items tend to produce stronger as well as more new responses than either small-weak or large-weak items. Such results are consistent with the possibility that the single response procedure provides a good estimate of the

strongest but not the weakest associates, but the chaining alternative must be ruled out before accepting this conclusion.

Insert Table 3 about here

To evaluate chaining we assumed that the first response could potentially influence the second response only if it was related to it. If 15 new words were added as second responses for a given normed item and the first response was associated to each of them according to the single-response norms, then 100% of the new words could have been produced as a result of chaining. However, this calculation indicated that the chaining potential was 26%, 20% and 8%, respectively, for small-strong, small-weak, and large-weak items. This potential seems small and does not map into the pattern of new responses for these conditions as shown in Table 3. These results do not rule out chaining in continuous responding tasks nor do they rule out chaining based on the combination of cue and first response, but they do suggest that the great majority of the new words produced here were probably not the result of response-to-response chaining. Single response norms seem to provide a good estimate of the strongest but not the weakest associates in the associative set of a word.

Evaluating the Hypotheses

The absolute strength hypothesis. The absolute strength hypothesis predicts that the ratios of first to second response strength should approach 1.0 for the primary associates, secondary associates, and so on, for each associative rank and item-type. For example, if *money* has a probability of .75 as a first response it is predicted to have a (conditional)

probability of .75 as a second response. These ratios were calculated for each item-type for the first six associative ranks, and this information is displayed in Figure 1. As can be seen, the ratios varied substantially as a function of both item-type and rank and, for the most part, they deviated from the expected ratio of 1.0. The ratios for small-strong, small-weak and large-weak items were, respectively, .85, 1.19, and 1.59. Item-type was a significant source of variance, $F(2, 61) = 24.65$, MSe = .66, and an LSD = .15 indicated that all means were significantly different. The effects of associative rank were significant, $F(5, 305) = 14.84$, MSe = .76, LSD = .22, as was the interaction between item-type and rank, $F(10, 305) = 1.90$. As shown in the figure, the strength ratios were considerably above 1.0 for the primary associate (Rank 1) for all item-types, but particularly for small-strong items. These ratios drop substantially from the primary to the secondary associate, settle into different levels for each item-type, and then gradually decline.

Insert Figure 1 about here

These findings indicate that ratios of first to second response strength are not constant across ranks and moreover they vary with item-type. This shortcoming was most evident for the primary associate, particularly for small-strong items which have highly skewed response distributions. Even when the primary associate was excluded, the strengths of the second responses tended to be either underestimated (small-strong) or overestimated (large-weak) by the strengths of the first responses. Such results are inconsistent with expectations derived from the absolute strength hypothesis and it is difficult to understand how a weaker version of this

hypothesis could explain the findings. Put simply, the index of strength for a given word changes substantially when computed from second as compared to first responses, and such changes are inconsistent with the supposition that probability of response production *is* strength.

The relative strength hypothesis. According to the relative strength hypothesis, probability of response production is a *manifestation* of strength. This hypothesis assumes that reading a to-be-normed word activates a pool of related associates in the informational domain specified by the instructions, e.g., meaning or rhyme, and then the strongest associates in this pool are selected and reported. As illustrated in Figure 2 for the associates of the word BRUSH, the relative strength hypothesis assumes that the strength of any given response is not given by a single value but by a distribution of values. This hypothesis was implemented in a simulation model[1] assuming that variability in sampling occurs both within and between participants (evidence for such variability can be found in Simpson and Voss (1967)). Within-participant variance occurs because of recent experience, e.g., a recent accident could increase the probability of producing *death* to BRUSH. Between-participant variance occurs because of life experience, e.g., being raised in the country could increase the probability of producing *fire* to BRUSH.

For reasons of simplicity and identifiability, the model assumed that the standard deviations of the strength distributions underlying each response were equal to 1.0. As also shown in Figure 2, the model establishes a response criterion. The response criterion was imposed because participants sometimes failed to respond or responded with unreliable idiosyncratic responses (Nelson & Schreiber, 1992). When all strengths were less than the criterion, the model assumed that either a non-response or an idiosyncratic response occurred.

Because the response probabilities depend on the differences in the means of these distributions rather than on the values of the means themselves, the response criterion could be set at 0.00 without loss of generality. Establishing the response criterion represented an important aspect of the model because it created a point of reference whereby the means of the strength distributions belonging to different words could be meaningfully contrasted.

Insert Figure 2 about here

The first step in simulating the second response probabilities was to estimate the means of the strength distributions by using the response probabilities from the first response on an item-by-item basis. For instance, in the first-response position, BRUSH has 6 associates, including *hair* (.447), *tooth* (.247), *comb* (.147), *paint* (.080), *fire* (.033), and *death* (.013). The summed probability of non-responses and idiosyncratic responses was .033. These probabilities served as constraints on the output of the model and an optimization procedure was used to estimate the means of the strength distributions. For BRUSH, the estimated mean strengths were *hair* (.76), *tooth* (.29), *comb* (-.06), *paint* (-.42), *fire* (-.87), and *death* (-1.27).

The second step in simulating the model used the estimated mean strengths to predict the second-response probabilities. A single strength was sampled from each of the strength distributions underlying each response to every normed word. For example, a single strength was sampled for hair, tooth, and so on, for the normed word BRUSH. The model assumed that the associate with the highest sampled strength for a given person and moment was

produced as the first response, and that the associate with the second highest sampled strength was produced as the second response. Only a single sample of strengths was taken from each of the response distributions, and the same sample was used in determining both the first and the second responses. If on either response occasion the strength fell below the strength criterion, it was assumed that the participant failed to respond or produced an idiosyncratic response. As in the experiment, the model could not produce the same response twice.

Because only a single strength was sampled from each response distribution, the model captured the notion that item selection can affect the probability of the second response. For a given participant at a given moment, the primary associate may not be selected as the first response because a weaker strength within its distribution was sampled. If a weaker strength was sampled, then the probability of selecting that response as the second item will also be lower. For example, if a weak strength underlying *hair* was sampled when given BRUSH as the cue, then *tooth* or some other response may be selected in preference. Such selection should be most apparent for the primary associate. The probability of selecting the primary associate for the second output position is determined by participants who did not produce it as the first response. In such cases, it is likely that such participants sampled a weak strength from the primary's distribution. The probability that these participants will produce this item as a second response will also be lower because stronger strengths are more likely to have been sampled from other responses. Item selection is exhibited by the model without introducing parameters requiring optimization.

The simulation results generated by the relative probability model along with the first response data and the second response data are displayed in Figure 3 for each item-type. The

critical comparisons for the model involved the second response probabilities for each associative rank, and as can be seen, the model fit the data exceptionally well, with a single exception. For the small-strong items, the model underestimated the relative probability of the primary associate. An item-type by rank by fit (model versus data) analysis of variance produced two reliable sources of variance involving fit, Item-Type x Fit, $F(2, 117) = 4.98$, $MSe = .002$, and Item-Type x Rank x Fit, $F(10, 585) = 2.69$, $MSe = .003$, $LSD = .024$. The significance of these sources indicated that the model significantly failed only in underestimating the probability of the primary associate for small-strong items. The model predicted a probability of .33 but the second response data showed a probability of .39. None of the remaining 17 contrasts between the model and the data were reliably different. Despite the absence of free parameters, the relative strength model provided a good fit to the second response probability data. The single exception involving the primary for small-strong items was probably the result of setting the standard deviation for all item strengths to one. Presumably, for words with strong primaries, the standard deviation is likely to be less than one.

Insert Figure 3 about here

General Discussion

Measurement issues. Free association produces a hierarchical distribution of responses that vary in probability that may or may not map directly onto the associative structure of a single individual (Bilodeau & Howell, 1968; Bruder, 1968; Fox, 1968; Nunnally,

Koplin, Blanton & Shaw, 1967; Osipow & Grooms, 1965; Simpson & Voss, 1967). This distribution constitutes an array of responses for a population of individuals and, as with other norms, free association reveals commonality within a social unit that arises as a result of similar experience. Such experience creates shared associative structures and free association is but one method that can be used to manifest and index these structures in ways that allow their manipulation in research. Although such measurements may or may not be representative of the associative structure of a single individual, such norms are rarely used in this way given the early failures in using free association for clinical diagnosis. Instead, measures of pre-existing experience made on one group of participants effectively predict the performance of similar groups of participants in related tasks. Free association has been used to provide an index of the strengths of forward, backward, mediating and shared connections between pairs of words (e.g., Nelson et al., 1998). It has also been used to index the size (Howe, 1972; Nelson et al., 1998) and connectivity of these structures (Deese, 1965; Nelson et al., 1998). The success of these indices in predicting free recall (Deese, 1965), cued recall and recognition (Nelson et al., 1998), as well as false memories (Deese, 1959; McEvoy, Nelson, & Komatsu, in press) indicates that they effectively capture key aspects of pre-existing lexical experience shared by free association and these tasks. Such success depends on capturing the aspects of language experience likely to be common to both free association and the criterial task.

Such findings indicate that free association norms provide a valid means for predicting performance under a variety of conditions. This conclusion is not in contention and has been widely accepted for years (Cramer, 1968; Deese, 1965). Primarily as a result of this success researchers have extended the free association procedure to other domains besides meaning,

including rhyme (Nelson & McEvoy, 1979), stem completion (Graf & Williams, 1987; Nelson, Canas, Bajo & Keelean, 1987), and picture naming (McEvoy, 1988). These efforts reflected an interest in controlling and manipulating the strengths of various types of cues in both explicit and implicit memory tasks, but in experiencing success with such indices, no one seems to have been overly concerned with their reliability. Such indifference is understandable because reliability serves to set limits on validity and, when a predictor is valid, it is reasonable to assume that both free association and task performance are reliable or, at the least, reliable enough. Nevertheless, more precise evaluations require more precise measures of item reliability in free association and in the criterial task. For example, the reliability of each procedure must be determined in order to determine how well free association indices predict cued recall at the item level.

The results of the present experiments suggest that discrete free association provides reliable indices of both response probability and set size, at least for the strongest associates. When based on first response to mind, the overall reliability of the probability of response index was .89 for the 120 words used in the present sample which agrees well with the results of earlier studies (Jenkins & Palermo, 1965; Nelson & Schreiber, 1992). Similarly, set size was also a reliable characteristic, averaging .85 for these and other words. Although these estimates are based on few words relative to the total number, they suggest that free association probability and set size appear to represent stable characteristics across different samples of individuals. Nevertheless, the present study revealed some differences in the probability index for items having different distributional characteristics. Words with small sets of associates and strong primaries showed the highest reliability, followed by words with small sets of associates

and weak primaries, and those with the largest associative arrays showed the lowest reliability. When the computation of response probability is determined by the first response produced, reliability is generally high but it is somewhat higher for words with fewer and/or stronger primary associates.

When the basis for computing reliability consisted of first and second responses in the dual-response norms, the reliabilities declined for both probability and set size indices. For each measure, the decline was most evident for words with smaller sets of associates, particularly for those words with stronger primaries. For the probability index, the decline in reliability occurred because the probability of a given second response tended to be lower than its probability as a first response. This drop was most evident for the primary associate of each word-type and was more evident for small-strong items than for the other items. This drop in probability was accompanied by a corresponding increase in set size. For the set size index, the decline in reliability occurred because set size increased by about 16 words for small-strong items and by about 11-12 words for the other item-types. The results of a "chaining" analysis suggested that most of these new items were probably not produced as responses to the first response. Virtually all of these new items could be defined as weak associates of the normed word typically not produced when only a single response is collected for each normed word.

Taken together, these findings suggest several conclusions about free association. The main conclusion is that the relatively high reliability of both strength and set size indices for first responses indicates that this procedure captures stable aspects of associative networks, aspects that can be useful in predicting performance in a variety of tasks. However, the lower reliabilities for both indices when the estimates are based on first and second responses for the

same sample of participants places important limits on this conclusion. Allowing a second response and, by implication continuous responding, reduces the reliability of both strength and set size indices. Allowing a second and presumably any additional response tends to underestimate the strength of the primary associate, more so for some item-types than for others. People who do not produce the primary associate as the first response tend to produce it as a second response at a *lower* probability. This finding suggests that free association can be influenced by item selection resulting from variation that arises within or between individuals or both.

A second limit is that using only the first response data underestimates the number of associates connected to a given word. When offered the opportunity for a second response about a dozen new items are produced. These items are generated by 2-3 people each time so that such items cannot be considered as unreliable as are idiosyncratic responses (Nelson & Schreiber, 1992). Such associates seem to be weak members of the associative set. These findings indicate that discrete first-response free association provides a reliable index of the strongest but not the weakest associates in the set. The procedure merely identifies the strongest members of the set in relative order of strength, and it also reliably indicates how many strong associates there are. Free association should not be relied upon for identifying all of the associates in a given set, at least with sample sizes hovering around 150 participants. Identifying all associates may be possible with sample sizes approaching 1,000 participants (e.g., Jenkins & Palermo, 1965), but using such large sample sizes is unfeasible because it comes at the cost of norming more items. Having a large pool of normed items is essential when using norms to

study the influence of pre-existing connections on recall (e.g., Deese, 1965; Nelson, Schreiber, & McEvoy, 1992; Nelson et al., 1998).

Having acceptable indices of the strongest associates may be all that is necessary for many types of research. In cued recall, for example, test cues are used to prompt the recall of their studied target words. Cues with probabilities of .05 or less are likely to be ineffective and cues stronger than .50 are likely to produce ceiling effects. An examination of 2,131 pairs of words used in our past cued recall work indicated that, although a wide range of cue-to-target strengths have been used (.01-.82), strength averaged .17 (SD = .12). Approximately two-thirds of the strengths ranged from .04 to .29, and such values were generally selected to produce recall performance that was above the floor but off the ceiling so the influence of other variables could be investigated. Finally, although true set size may be underestimated by single-response norms, both cue and target set size effects are robust phenomena (Nelson, Schreiber, & McEvoy, 1992). This could not be the case if the discrete free association task did not reliably index the set size of the strongest associates. In the studies mentioned above, small-strong items were generally avoided in these studies because their effectiveness as test cues and their recall as targets falls between small-weak and large-weak items (see Nelson & Bajo, 1985). The present findings suggest that one explanation for such intermediate performance is that small-strong items may have medium set sizes that fall between those of small-weak and large-weak items

Theoretical issues. The results of the dual-response norms were inconsistent with expectations derived from the absolute strength hypothesis and therefore with the hypothesis that probability of free association *is* strength. Rather than being constant as predicted, the

ratio of first-to-second response strength differed substantially as a function of associative rank and for words having different response distributions. Such results are inconsistent with expectations based on the absolute strength hypothesis and with the idea that strength represents a single abstract entity that increases in potency with experience. Free association provides a reliable index of strength but it is not absolute strength that is being measured. It is incorrect to say that *table* is 2.5 times stronger than *sit* as a response to the cue CHAIR when the probabilities of free associating the responses *table* and *sit* are, respectively, .50 and .20. These probabilities estimate mean strengths with unknown amounts of error variance and an unknown underlying measurement scale. Assuming that an ordinal scale of measurement has been achieved, all that can be said with some assurance is that *table* is a reliably stronger associate of CHAIR than *sit*.

The relative strength hypothesis faired much better in explaining the present findings. According to this interpretation, probability of response production is a manifestation of strength, not strength itself. This hypothesis is compatible with the assumption that the strength of a response reflects the number of its instances in memory, with stronger associations reflecting larger numbers of instances (Hintzman, 1976; Jacoby & Brooks, 1984; Landauer, 1975). Reading a word presumably activates a pool of related associates in the informational domain specified in the test instructions (e.g., meaning), and then the strongest associates in this pool are selected and reported. The choice of any response in the set is based on its mean strength relative to the mean strengths of the other associates in the set. The relative strength hypothesis assumes that the strength of each response can be represented as a distribution of values. This hypothesis was implemented in a simulation model which assumed that variability in sampling

occurs both within- and between-participants. For a given individual, strength can vary from moment-to-moment, and for different participants the same response can be represented at different strengths depending upon their experience. For example, those prone to substance abuse should be prone to differentially selecting meanings of drug related words that reflect their experience (e.g., Stacy, 1997).

The simulation sampled a single strength for each response of a normed word, and assumed that the associate with the highest and second highest strengths were produced, respectively, as first and second responses in the dual-response norms. If on either response occasion the strength fell below a criterion, the model assumed that the participant failed to respond or produced an idiosyncratic response. Because only a single strength was sampled from each response distribution, the model can explain how the probability of a given response can be lower when produced on the second as opposed to the first production opportunity. A stronger associate in the set is not as likely to be selected as a response whenever a weaker strength within its distribution has been sampled. The probability of selecting the strongest associate on the second response opportunity is determined by participants who did not report this item as the first response. According to the model, the probability that this item will be produced as a second response will be lower because stronger strengths from other responses in the set presumably have been sampled. The model exhibits item selection without having to set and estimate specific parameters, and with only a single exception, it provided a good fit to the second response probability data in 17/18 comparisons. The exception occurred for the primary associates of small-strong words and was the result of setting a constant standard deviation for all items that presumably was too large for such words.

The activation and sampling processes associated with the relative strength hypothesis imply that free association should be a good predictor of performance whenever the criterion task involves similar representations and/or processes. Predictive utility should hold even though the model assumes that probability of free association provides a manifestation of strength rather than strength itself. What matters more in prediction is process and representational similarity, not whether the underlying measurement scale is relative or absolute. Within the limits imposed by reliability, the strength of a measured association should be and is a good predictor of lexical decision (e.g., Canas, 1990), naming (McEvoy, 1988), free recall (Jenkins & Russell, 1952; Deese, 1959), and cued recall (e.g., Bahrick, 1970; Nelson & McEvoy, 1979). It is a good predictor because such tasks allow or require people to use some words to judge or to recall related words. Similar representations and processes are involved in free association and in each of these tasks, but recognizing this similarity is not tantamount to claiming that no processes differentiate these tasks (e.g., see Gillund & Shiffrin, 1984; Humphreys, Wiles, & Dennis, 1994). They share some characteristics and differ on others. For example, although both free association and cued recall are instigated by presenting cues, the encoding conditions and the goal of each task are different. In free association, the goal is to produce any word that is related to the cue without reference to a recent study experience, whereas the goal in cued recall is to produce a related word that is a member of a recently studied list. This encoding-goals difference limits the power of free association measures as predictors of cued recall performance but it does not preclude prediction based on common representations and processes.

By way of final comment on these two hypotheses, we note that critics might argue that only a few at most have ever held the absolute strength view of association norms. Explicit references to this position in the literature are admittedly rare (e.g., Rosen & Russell, 1952 argue for it and Gillund & Shiffrin, 1984 argue against it). Nevertheless, the absolute position seems to be implicit in the early empirical work in the field. Moreover, in contemporary work whenever researchers manipulate two or more levels of pre-existing strength for example, they put items with similar strengths into the same condition (e.g., Nelson & McEvoy, 1979). In so doing, they are operating as if association probability is equivalent to association strength rather than a manifestation of such strength. Because of variability association probability and association strength are not necessarily the same in a relative strength model, and the success of such manipulations hinges on both the effects of strength in the task and on the reliability of its measurement in free association. The results of the present paper suggest that the success of past manipulations of association probability may owe much to the reliability of the strength index.

Author's Note

This research was supported by grants MH16360 from the National Institute of Mental Health to D. L. Nelson and AG13973 from the National Institute on Aging to C. L. McEvoy.

Correspondence concerning this article should be addressed to Douglas L. Nelson, Department of Psychology, University of South Florida, Tampa, FL, 33620-8200.

nelson@luna.cas.usf.edu

Footnotes

[1] The simulation model was developed by Simon Dennis alone, and the details of the simulation and an example are available on request.

References

Anderson, J. R., & Bower, G. H. (1974). *Human Associative Memory*. Washington, DC: Hemisphere Publishing.

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review, 77,* 215-222.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph, 80,* 1-46.

Bilodeau, E. A., & Howell, D. C. (1968). Stimulated recall of idiosyncratic and cultural free associates. *Journal of Verbal Learning and Verbal Behavior, 7,* 348-350.

Brown, A. S. (1976). Catalog of scaled verbal material. *Memory & Cognition, 4,* 1S-45S.

Bruder, G. A. (1968). Comparisons of individual and cultural response hierarchies based on continued association. *Journal of Verbal Learning and Verbal Behavior,* 7, 1119-1121.

Canas, J. J. (1990). Associative strength effects in the lexical decision task. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 42,* 121-145.

Cieutat, V. J. (1963). Associative indices for 446 randomly selected English monosyllables, bisyllables, and trisyllables. *Journal of Verbal Learning and Verbal Behavior, 2,* 176-185.

Cramer, P. (1968). *Word Association.* New York: Academic Press.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore, MD: The Johns Hopkins Press.

Fox, P. W. (1968). Recall and misrecall as a function of cultural and individual word association habits and regulation of the recall environment. *Journal of Verbal Learning and Verbal Behavior, 7*, 632-637.

Galton, F. (1880). Psychometric experiments. *Brain, 2*, 149-162.

Gaeth, J. H., & Allen, D. V. (1966). Associative values for selected trigrams with children. *Journal of Verbal Learning and Verbal Behavior, 5*, 473-477.

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

Graf, P., & Mandler, G. (1984). Activation makes words more accessible, but not necessarily more retrievable. Journal of Verbal Learning and Verbal Behavior, 23, 553-568.

Graf, P., & Williams, D. (1987). Completion norms for 40 three-letter word stems. *Behavioral Research Methods, Instruments & Computers, 19*, 422-445.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 47-91). New York: Academic Press.

Howe, E. S. (1972). Number of different free associates: A general measure of associative meaningfulness. *Journal of Verbal Learning and Verbal Behavior, 11*, 18-28.

Humphreys, M. S., Wiles, J., & Dennis, S. (1994). Towards a theory of human memory: Data structures and access processes. *Behavioral and Brain Sciences, 17*, 655-666.

Jenkins, J. J., & Palermo, D. S. (1965). Further data on changes in word-association norms. *Journal of Personality and Social Psychology, 1*, 303-309.

Jenkins, J. J., & Russell, W. A. (1952). Associative clustering in recall. *Journal of Abnormal and Social Psychology, 47*, 818-821.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 1-47). New York: Academic Press.

Johnson, R. E. (1968). The reliability of association norms. *Journal of Verbal Learning and Verbal Behavior*, 7, 1054-1059.

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

Landauer, T. K. (1975). Memory without organization: Properties of a model with random storage and unidirected retrieval. *Cognitive Psychology, 7*, 495-531.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and the representation of knowledge. *Psychological Review, 104*, 211-240.

McEvoy, C. L. (1988). Automatic and strategic processes in picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 618-626.

McEvoy, C. L., & Nelson, D. L. (1982). Category name and instance norms for 106 categories of various sizes. *American Journal of Psychology, 95*, 581-634.

McEvoy, C. L., & Nelson, D. L., & Komatsu, T. (in press). What is the connection between true and false memories? The differential roles of interitem associations in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Nelson, D. L., & Bajo, M. T. (1985). Prior knowledge and cued recall: Category size and dominance. *American Psychologist, 98*, 503-517.

Nelson, D. L., & Brooks, D. H. (1974). Relative effectiveness of rhymes and synonyms as retrieval cues. *Journal of Experimental Psychology, 102*, 503-507.

Nelson, D. L., Canas, J., Bajo, M. T., & Keelean, P. (1987). Comparing word fragment completion and cued recall with letter cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 542-552.

Nelson, D. L., & McEvoy, C. L. (1979). Encoding context and set size. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 292-314.

Nelson, D. L., & McEvoy, C. L. (in press). What is thing called frequency? *Memory & Cognition*.

Nelson, D. L., & McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme and word fragment norms. http://www.usf.edu/FreeAssociation/.

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review, 105*, 299-324.

Nelson, D. L., & Schreiber, T. A. (1992). Word concreteness and word structure as independent determinants of recall. *Journal of Memory and Language, 31*, 237-260.

Nelson, D. L., Schreiber, T. A., & Holley, P. E. (1992). The retrieval of controlled and automatic aspects of meaning on direct and indirect tests. *Memory & Cognition, 20*, 671-684.

Nelson, D. L., Schreiber, T. A., & McEvoy, C. L. (1992). Processing implicit and explicit representations. *Psychological Review, 99*, 322-348.

Noble, C. E. (1952). An analysis of meaning. *Psychological Review, 59*,421-430.

Nunnally, J. C., Koplin, J. M., Blanton, R. L., Shaw, R. (1967). Individual differences in word association in relation to paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 107-111.

Osipow, S. H., & Grooms, R. R. (1965). Comparisons between cultural and individual associative response hierarchies. *Journal of Verbal Learning and Verbal Behavior, 4*, 94-97.

Palermo, D. S., & Jenkins, J. J. (1964). Word association norms: Fourth grade through college. Minneapolis, Minnesota: University of Minnesota Press.

Pincus, T., Pearce, S., McClelland, A. Farley, S., & Vogel, S. (1994). *Journal of Psychosomatic Research, 38*, 347-353.

Postman, L. & Keppel, G. (1970) *Norms of Word Association*. New York: Academic Press.

Roediger, H. L., III. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*, 1043-1056.

Rosen, E., & Russell, W. A. (1957). Frequency-characteristics of successive word association. *American Journal of Psychology, 70*, 120-122.

Simpson, P., & Voss, J. F. (1967). Stability of response hierarchies. *Journal of Experimental Psychology, 75*, 170-174.

Stacy, A. W. (1997). Memory activation and expectancy as prospective predictors of alcohol and marijuana use. *Journal of Abnormal Psychology, 106*, 61-73.

Watkins, P. C., Vache, K., Verney, S. P., & Mathews, A. (1996). Unconscious mood-congruent memory bias is depression. *Journal of Abnormal Psychology, 105*, 34-41.

Weldon, M. S., Coyote, K. C. (1996). Failure to find the picture superiority effect in implicit conceptual memory tests. . *Journal of Experimental Psychology: Human Learning and Memory, 22,* 670-686.

Appendix A

The words and their reliability's

Small-Strong	Reliability	Small-Weak	Reliability	Large-Weak	Reliability
AIRPORT	1.00	ARTERY	.92	ACCIDENT	.96
ARITHMETIC	1.00	BIBLE	.96	AGE	.98
ASTRONOMY	1.00	BRUSH	.98	AIRPLANE	.97
AUNT	1.00	CALENDAR	.72	ANIMAL	.97
BANK	1.00	CAMERA	.97	BEAD	.99
BLAZE	1.00	CANDLE	.95	BLOOD	.98
BRIDE	.99	CASKET	.99	BREAST	.81
CASH	1.00	COFFIN	.95	BROWN	.35
CASHEW	1.00	CORRIDOR	1.00	BURN	.88
COBRA	1.00	DARK	.79	CAPE	.40
CRIB	1.00	EMERALD	.88	CARD	.52
DAY	.95	EMPLOYER	.92	CHISEL	.94
DILL	1.00	EVENING	.95	COOKING	.94
EXAM	.99	EXIT	.66	CURVE	.31
HAMMER	1.00	FRAGRANCE	1.00	DANCER	.98
HIVE	1.00	GLUE	.99	DECISION	.72
HUSBAND	1.00	HUMOR	1.00	DIFFERENCE	.98
ILL	1.00	LIBERTY	.92	DRAGON	.96
INFANT	1.00	MALL	.38	ENTERTAIN	.21
JOG	.99	MELODY	1.00	FROG	.61
KEG	1.00	MINUTE	.92	INSTINCT	.94
KITTEN	.99	MONK	.87	IRON	.56
LAMP	.95	NEEDLE	.85	JEWEL	.98
LOST	1.00	PETROLEUM	.92	LEADER	.97
MARGARINE	1.00	POND	.77	MAIN	.97
OMELET	1.00	PUDDLE	.87	MANNER	.48
OPENER	.99	PUMPKIN	.82	MILDEW	.79
PHYSICIAN	1.00	RAZOR	.99	MISCHIEF	.99
PONY	.98	SABER	.42	PARADISE	.95
QUESTION	.99	SHAMPOO	.78	PLASTIC	.63
QUIZ	1.00	SHEET	.86	PUNCH	.91
SADDLE	1.00	SHERIFF	.94	RACCOON	.97

Appendix A (continued)

Small-Strong	Reliability	Small-Weak	Reliability	Large-Weak	Reliability
SALTINE	.99	SHOELACE	.91	RAW	.94
SCISSORS	1.00	SIMILAR	.86	REWARD	.99
THREAD	.97	TAVERN	.82	ROAD	.89
THUNDER	.97	THIRST	.65	SCULPTURE	.96
TOAD	1.00	THUMB	.73	SHOULDER	.84
TULIP	1.00	TWIG	.96	SQUASH	.96
TUNA	1.00	VIOLET	.95	TIP	.76
YOLK	1.00	WRIST	.92	TRAIL	.68

Table 1

Mean response probabilities for the first six associates of the first and second responses for each item-type.

		Rank of the First Six Associates					
Item-Type	Response Order	1	2	3	4	5	6
Small Set-Strong Primary	First Response	.73	.07	.04	.02	.02	.01
	Second Response	.39	.13	.08	.06	.05	.03
Small Set-Weak Primary	First Response	.39	.21	.12	.07	.04	.03
	Second Response	.26	.22	.13	.10	.05	.03
Large Set-Weak Primary	First Response	.31	.12	.08	.06	.04	.04
	Second Response	.19	.09	.07	.05	.03	.03
Pooled Over Item-Type	First Response	.48	.13	.08	.05	.03	.02
	Second Response	.27	.15	.09	.07	.04	.03

Table 2

Set size as a function of item-type for the single- and dual-response norms (standard deviations).

Item-Type	Single-Response	Dual-Response Norms		
		First Response	Second Response	Pooled Responses
Small Set-Strong Primary	5.12	6.43	19.08	20.20
	(1.26)	(2.93)	(3.71)	(3.71)
Small Set-Weak Primary	7.32	9.33	15.18	17.32
	(1.02)	(2.55)	(2.63)	(3.53)
Large Set-Weak Primary	21.33	17.15	21.85	26.68
	(2.12)	(2.85)	(3.69)	(4.22)

Table 3

Mean strength and number of new words appearing in the dual-response norms that did not appear in the single-response norms as a function of item type (standard deviations).

Item-Type	Mean Strength of New Words	Mean Number of New Words
Small Set-Strong Primary	.016	15.65
	(.004)	(4.14)
Small Set-Weak Primary	.013	10.90
	(.004)	(3.28)
Large Set-Weak Primary	.013	12.38
	(.004)	(5.19)
Pooled Over Item-Type	.014	12.98
	(.004)	(4.68)

Figure Captions

Figure 1. Ratio of first response probability to second response probability as a function of item-type and associative rank.

Figure 2. Hypothetical strength distributions underlying the responses for the word BRUSH.

Figure 3. Probability of response as a function of item-type and associative rank for the second response data and the model. The first response data (also shown) was used by the model to estimate second response probability.