

Item-specific effects in recognition failure: Reasons for rejection of the Tulving–Wiseman function

ARILD LIAN

University of Oslo, Oslo, Norway

ARNOLD L. GLASS

Rutgers University, New Brunswick, New Jersey

and

RUTH K. RAANAAS

University of Oslo, Oslo, Norway

The present paper addresses the problems of whether recognition failure of recallable words is a function of both recognition and recall, and whether recognition failure is restricted to a small and specifiable subset of study items. A meta-analysis of the Nilsson–Gardiner database (Nilsson & Gardiner, 1993) showed that recognition given recall was positively correlated with recognition and negatively correlated with recall. Two new experiments are reported, the first one using 48 word pairs for which recognition failure was found in previous studies. An item analysis of the data demonstrated that recognition failure occurred primarily with noun–adjective pairs. The second experiment compared Norwegian–American and American–Norwegian name pairs. Wide deviation from the Tulving–Wiseman function (Tulving & Wiseman, 1975) was observed for the latter condition. In both conditions, recognition failure occurred with only the items for which the beginnings of names shared three or more letters. It is concluded that recognition failure occurs when there exists a relationship between the members of an A–B pair that is independent of their pairing in the study context. The Tulving–Wiseman function is the result of collapsing across items in the analysis of previous studies.

Both remembering and forgetting are imperfect phenomena. A list item that is forgotten on one occasion may be remembered on a later occasion. Tulving and his colleagues (Tulving, 1983) have contributed many studies of such fluctuations in retrieval that make use of a common experimental paradigm, a procedure that has been called the *recognition-failure paradigm*. In this paradigm, subjects study a list of paired, A–B, items (e.g., *cabbage round*). After a certain retention interval, they are given a recognition test for the B words (e.g., *round*) in the absence of the A words. This is followed by a cued recall test in which they are given A (e.g., *cabbage*) and asked to retrieve B. Tulving found that a certain proportion of the B words that were not recognized when used as targets on the recognition test were nevertheless recalled in response to their corresponding A word in the cued recall test. This finding is commonly referred to as recognition failure of recallable words, or just *recognition failure*.

Tulving (1983) organized the results of experiments using the recognition-failure paradigm in 2×2 contingency tables categorized according to success and failure in recognition (RN) and success and failure in recall (RC). Such tables give 4 unconditional and 8 conditional probabilities. The relation between recognition and recall is generally expressed by the conditional probability of recognition given recall, $P(RN|RC)$. This measure, which is also called *recognition success*, is of course the complement of $P(nRN|RC)$ (i.e., recognition failure of recallable words). Tulving and Wiseman (1975) reported that recognition success was related to the unconditional probability of recognition, $P(RN)$. On the basis of a meta-analysis of the results of 40 conditions from 12 experiments using the recognition-failure paradigm, Tulving and Wiseman described the relationship by the following equation:

$$P(RN|RC) = P(RN) + .5[P(RN) - P(RN)^2]. \quad (1)$$

Subsequent research expanded the database for this function to over 300 studies (Nilsson & Gardiner, 1993). When data in the Nilsson–Gardiner (N–G) database were plotted in the Tulving–Wiseman (TW) graph space (i.e., with recognition success as the y-ordinate and probability of recognition as the x-abcissa), 92% of the data points conformed to the function. Therefore, some researchers

Preparation of this article was supported in part by the Norwegian Research Council during A. Lian's sabbatical at Rutgers University, 1994–1995. Correspondence should be addressed to A. Lian, Institute of Psychology, Box 1094, University of Oslo, N-0317 Oslo, Norway (e-mail: arild.lian@psykologi.uio.no).

—Accepted by previous editor, Geoffrey R. Loftus

(Cohen, 1985; Jones, 1984; Nilsson, Law, & Tulving, 1988) have treated the TW function as an empirical law. Many other researchers (e.g., Årlemalm, 1996; Flexser & Tulving, 1978; Gardiner & Nilsson, 1993; Tulving, 1983) relied on the meta-analysis of either the TW or the N-G database to claim that recognition is the *sole* determinant of recognition success. In other words, the assumption is that recognition success is not significantly correlated with the probability of recall.

Before proceeding to a meta-analysis of the N-G database, we will comment on Nilsson and Gardiner's (1993) criterion of agreement with the TW function. To distinguish between natural variations of recognition success and "deviations that can be regarded as abnormally large and falling outside the normal range" (p. 398), they proposed a critical ratio (*CR*) measure, which is the ratio of a single difference score (i.e., letting RS = recognition success, $DevTW$ = observed RS - predicted RS) and the standard deviation for all these scores in the database. For a confidence interval of 90%, the *CR* is 1.64, which is reached by a $DevTW$ of .1312. A total of 92% of the $DevTW$ s in the N-G database were below this value and, hence, were said to agree with the TW function.

If we use a more conservative method, which creates narrower confidence intervals, then of course more data points are outside of their bounds. For example, Muter (1978), by way of a normal approximation to the binomial distribution, defined the significance of a difference score as $z = DevTW / \{[\text{predicted } RS (1 - \text{predicted } RS)]/N\}^{1/2}$, where N is the number of items recalled (see also Neely & Payne, 1983). According to this definition only 67% of the conditions in the N-G database yielded $DevTW$ s within a 90% confidence interval.

However, any test of the TW function that merely determines whether the observed data points fall within the confidence intervals of the points predicted by the function is ultimately less stringent than a test of the amount of variance that the TW function accounts for in the observed data points. A test of the fit of the TW function to the data points by itself is less interesting than a comparison of the amount of variance explained by the TW function with the amount explained by other plausible functions. Here, we will be most concerned with whether the TW function accounts for more variance than other functions that include recall.

The claim that the TW function is an accurate description of the relationship between recognition success and recognition has not escaped controversy. Hintzman (1992) and Riefer and Batchelder (1995) have argued that the TW function does not adequately describe the data at low levels of recall. Partly in response to this type of criticism, Gardiner and Nilsson (1990, 1993) have acknowledged that there are two categories of exceptions to the law—encoding exceptions and retrieval exceptions—which occur at low and high levels of recall, respectively. Encoding exceptions occur whenever the A terms are so

weakly encoded as to make the cued recall test functionally like one of free recall. Retrieval exceptions occur when the cued recall test, due to informational overlap between the A and B terms, has turned into one of "cued recognition" (Gardiner, 1994).

Hintzman's (1987, 1992) more recent specific objections were initially made more generally in Hintzman (1980). Hintzman (1980) showed that a contingent relation such as that shown by the TW function (i.e., that recognition success is a function of recognition), could appear in a subjectwise analysis of a data set even if it were true for a few or none of the individual items in the data set. Nevertheless, despite his more recent objections, Hintzman has not shown that the TW function holds for only some of the items in the typical recognition-failure experiment, which has allowed Hayman and Tulving (1989) to characterize these objections as merely theoretical.

The purpose of the present study was to further broaden both the sophistication of the data analysis and the database on which an understanding of recognition failure is based in order to determine its generality across items. To this end, we performed a new meta-analysis of the N-G database, as well as two new experiments. Experiment 1 used the same word pairs for which recognition failure has been previously demonstrated. In contrast, Experiment 2 used names similar to those for which recognition failure has not been found.

META-ANALYSIS OF THE NILSSON-GARDINER DATABASE

Consider the list of candidate variables for predictors of $P(RN|RC)$. The most obvious candidates are $P(RN)$ and $P(RC)$. An analysis of 300 cases in the N-G database (excluding cases 103 and 133, as suggested by Nilsson & Gardiner, 1993) shows that $P(RN|RC)$ correlates $r = .90$, confidence interval (CI) 95% = (.88, .92) with $P(RN)$, whereas the correlation between $P(RN|RC)$ and $P(RC)$ is a negligible $r = .06$, CI 95% = (-.05, .17). However, since $P(RN)$ was so strongly correlated with $P(RN|RC)$, it had the potential to act as a suppressor variable for a negative correlation between recall and recognition success. In fact, when we removed the correlation between $P(RN|RC)$ and $P(RN)$, there was a significant partial correlation between $P(RN|RC)$ and $P(RC)$ of $r = -.50$, CI 95% = (.41, .58). Furthermore, since Gardiner and Nilsson (1990, 1993) have argued that cases in which $P(RC)$ is low are actually encoding exceptions to the TW function, we performed an additional analysis in which cases of $P(RC) \leq .10$ were excluded. In this analysis, the partial correlation between $P(RC)$ and $P(RN|RC)$ was $r = -.56$, CI 95% = (.48, .63). $P(RN)$ and $P(RC)$ were used to predict $P(RN|RC)$ in a stepwise multiple regression analysis. A significance level of $p < .05$ was set for betas in this and the following regression analysis. In this analysis, $P(RN)$ and $P(RC)$ entered the equation with significant betas of .98 and $-.26$, respectively. When

$P(RC)$ entered the equation in Step 2, it added 6% to the explained variance. The final analysis yielded Equation 2 below, which accounted for 87% of the variance.

$$P(RN|RC) = .28 + .98P(RN) - .26P(RC). \quad (2)$$

Since the above analysis shows that both $P(RN)$ and $P(RC)$ are predictor variables of recognition success, estimated curves for data points relating $P(RN|RC)$ to $P(RN)$ will have different parameters depending on level of recall. We, therefore, made separate plots of $P(RN|RC)$ versus $P(RN)$ for two ranges of recall in the N-G database (again excluding cases 103 and 133). Figure 1A shows $P(RN|RC)$ plotted as a function of $P(RN)$ for $P(RC) \leq .40$. This graph is based on 77 conditions in the database. A total of 33, or 42.8%, of these cases yielded a Neely and Payne (1983) z above 1.65. Figure 1B shows the same function for $P(RC) > .40$. This graph is based on 223 conditions. A total of 29, or 13%, of them yielded a z above 1.65. In both graphs, we have estimated curves according to a linear and a quadratic model. Figure 1C shows the two sets of curves for $P(RC) \leq .40$ and $P(RC) > .40$ combined. The two sets of curves do not coincide. The same difference between the curve sets appears when cut-off points are chosen 1 standard deviation

(SD) below and 1 SD above the mean recall score. Notice that the linear curve for $P(RC) \leq .40$ has a higher intercept with the ordinate than does the corresponding curve for $P(RC) > .40$, indicating a higher dependence between recognition and recall for the lower range of recall. It appears from the Figures 1A and 1B that when the effect of $P(RC)$ is accounted for, a linear function may fit the data equally well. In fact, the correlation between the values predicted by Equation 2 and the observed recognition success was $r = .93$, $CI\ 95\% = (.91, .94)$, whereas the correlation between the values according to the TW function predicted by Equation 1 and the observed values was $r = .90$, $CI\ 95\% = (.87, .92)$.

Another way of examining the data is by plotting the absolute deviation from the TW function ($DevTW$) as a function of $P(RC)$. That is, the y -axis contains the difference between the value of $P(RN|RC)$ predicted from the TW function, Equation 1 above, and the obtained value. This plot is shown in Figure 2, which also includes an estimated curve relating the two variables. If recall does not affect the degree of conformity with the TW function, the curve would run parallel with the x -axis, and, granted perfect validity of the TW function, the curve would have

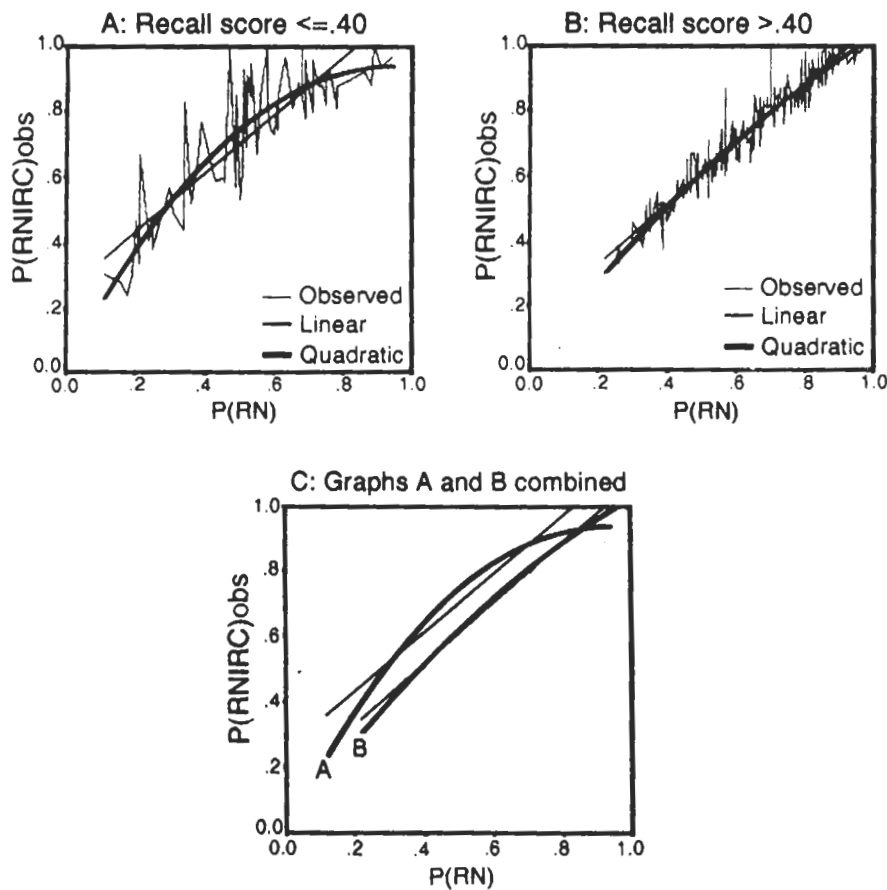


Figure 1. $P(RN|RC)$ as a function of $P(RN)$ for two levels of recall in the Nilsson-Gardiner database (Nilsson & Gardiner, 1993). Curves are estimated according to a linear and a quadratic model.

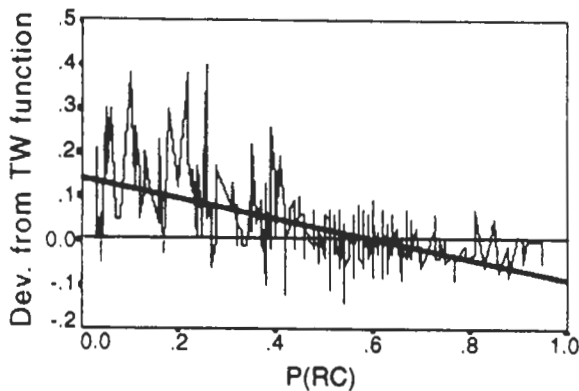


Figure 2. Linear curve estimation of deviation from the Tulving-Wiseman function plotted against $P(RC)$.

an intercept of 0. However, the plot shows that low levels of recall are associated with positive deviations from the TW function, whereas high levels of recall may be associated with negative deviations from this function, reflecting the failure of the TW function to take the negative correlation between $P(RC)$ and $P(RN|RC)$ into account.

However, before we are ready to reject the TW function as an adequate description of the data, we must acknowledge that the argument we have presented with respect to the role of $P(RC)$ in recognition-failure experiments is merely based on meta-analysis of the N-G database. It is constrained by the logic of interexperiment comparisons. So let us next examine the relationships among $P(RN|RC)$, $P(RN)$, and $P(RC)$ for data collected in a single experiment.

EXPERIMENT 1

The objective of Experiment 1 was to replicate a typical recognition-failure study in order to obtain data for our analysis. We, therefore, made use of the two lists of word pairs included in Tulving and Thomson's (1973) study and again in Watkins and Tulving's (1975) study. Both lists consisted of 24 weakly associated A-B terms. The 48 word pairs were translated into Norwegian with some minor adaptations to maintain a pattern of weak associations throughout the extended list.

Method

Subjects. Thirty-six students from the University of Oslo volunteered for participation in the experiment. The mean age of the subjects was 25 years (range = 20-30 years). All subjects were native speakers of Norwegian.

Materials and Procedure. Two study lists were created by randomizing the 48 pairs of words twice. Preexperimentally, 50 other subjects rated the degree of association between the words of a pair on a 3-point scale, with 1 for *weakly associated*, 2 for *somewhat associated*, and 3 for *strongly associated*. The English and Norwegian versions of the 48 word pairs and the mean association rating for each of these pairs are presented in Appendix A.

In each of the study lists, the A terms were written in lowercase and the B terms in uppercase letters. In the study task, the pairs were presented in equally sized columns on two pages. A test of recog-

nition was provided by mixing the B terms with 96 new words. Thus, 144 items were presented in 48 rows of 3. Each row contained one randomly positioned target word. The two distractors of each row (see Appendix A) were semantically related to the target. By randomly sequencing the rows twice, we created two versions of the recognition test.

Also, two versions of a cued recall test were created by arranging two random sequences of the A terms from the study list. Beside each A term was provided a space for the writing of the target word. The list of A terms was presented on two pages. The subjects were randomly assigned one of the two versions of the study list and the retrieval tests. They were given 8 min for the study task. Each subject was told that he/she would be tested later on his/her memory of the second member of each pair, the word written in uppercase letters. However, each subject was also told it was important to pay attention to the first member, since this word would help the subject to remember the last word of the pair. A retention interval of about 24 h was provided between the study task and tests. The next day, the subjects were first given the recognition test. They were told that each row contained one and only one target word from the study list. Hence, they were told to circle one word in each row and were asked to respond to every row of the test this way. The subjects were also told to rate their degree of confidence by writing 2 for *certain*, 1 for *maybe*, or 0 for *guessing* beside the circled word. All subjects had to work at least 10 min on the test, and they were encouraged to finish the test within 15 min. After an interval of 2-3 min "social-talk" subjects were given the cued recall test. They were told that the test included all A terms from the study list and that their task was to write beside each of these terms the one word that they thought appeared as the target in the study list. Again, they were told to rate their degree of confidence by writing 2 for *certain*, 1 for *maybe*, or 0 for *guessing* beside their written response. The same time limit that applied to the recognition test was also imposed on the cued recall test.

Results and Discussion

There were no significant differences between the two versions of the tests; therefore, the data from the two versions were analyzed together. Table 1 shows the mean scores for recognition, recall, and mean observed and predicted recognition success, together with the mean difference between the latter two scores. The data were tabulated both subjectwise (each data point represented the score of a single subject) and itemwise (in which each data point represented the score of a single item), so the *SDs* are listed separately for the subjectwise (SD_{SW}) and itemwise (SD_{IW}) tabulations.

The subjectwise tabulation of the data yielded a correlation of $r = .55$, $CI\ 95\% = (.27, .74)$ between $P(RN)$ and $P(RN|RC)$. $P(RC)$, however, had a nonsignificant correlation with $P(RN|RC)$ of $r = .27$, $CI\ 95\% = (.06, .55)$. A stepwise regression analysis based on subjectwise tabulation of the data was undertaken, with $P(RN|RC)$

Table 1
Retrieval Scores in Experiment 1

Variable	<i>M</i>	SD_{SW}	SD_{IW}
$P(RN)$	73	13	12
$P(RC)$	34	21	17
$P(RN RC)_{obs}$	84	12	14
$P(RN RC)_{TW}$	83	11	09
DevTW	01	11	11

as the dependent variable and with $P(\text{RN})$ and $P(\text{RC})$ as the independent variables. Only $P(\text{RN})$ entered the equation with a significant beta of .55. Equation 3 (below) explained 30% of the subject variance. Also, the values predicted by Equations 1 and 2 were used in another stepwise multiple regression analysis to predict $P(\text{RN}|\text{RC})$.¹ Only Equation 2 entered the regression equation with a significant beta of .53, and this variable alone explained 28% of the subject variance.

$$P(\text{RN}|\text{RC}) = .47 + .55 P(\text{RN}). \quad (3)$$

For each subject, the median confidence rating score was used to characterize his/her response criterion. The median rating scores for recognition and recall correlated [$r = .54$, $\text{CI } 95\% = (.25, .74)$]. Only the former score correlated with $P(\text{RN}|\text{RC})$ [$r = .37$, $\text{CI } 95\% = (.05, .62)$]. However, it did not enter the equation following $P(\text{RN})$ in a stepwise regression analysis where $P(\text{RN}|\text{RC})$ was the dependent variable.

The itemwise analysis was done by coding responses to each of the 48 items with reference to a contingency table for success or failure in recognition and success or failure in recall. Each cell was filled with the number of subjects who showed that contingent relationship between recognition and recall for that item. The results for each item are shown in Appendix B. This appendix shows that Item 16 (SOVN, SLEEP) is highly atypical, since none of the subjects was able to recall this target item. Because a recall score of 0 prevents an assessment of recognition success, this item was removed. The following analysis, therefore, is based on the remaining 47 items.

The itemwise tabulation of the data yielded a correlation of $r = .60$, $\text{CI } 95\% = (.35, .75)$ between $P(\text{RN})$ and $P(\text{RN}|\text{RC})$. $P(\text{RC})$ correlated with $P(\text{RN}|\text{RC})$ [$r = -.50$, $\text{CI } 95\% = (-.68, -.24)$] and with DevTW [$r = -.61$, $\text{CI } 95\% = (-.76, -.39)$]. In a stepwise regression analysis with $P(\text{RN}|\text{RC})$ as the dependent variable, both $P(\text{RN})$ and $P(\text{RC})$ entered the equation with significant betas of .58 and $-.49$, respectively. Equation 4, which accounts for 59% of the variance, summarizes the results of this analysis.

$$P(\text{RN}|\text{RC}) = .50 + .58 P(\text{RN}) - .49 P(\text{RC}). \quad (4)$$

When the predicted values of Equations 1 and 2 were used as independent variables to predict $P(\text{RN}|\text{RC})$, the result confirmed the corresponding regression analysis of the data set tabulated subjectwise (i.e., only the predicted value of Equation 2 entered the regression equation). This variable had a significant beta of .74 in an equation that explained 54% of the item variance.

The itemwise analysis confirms the relationship between recognition success and recall that first appeared in our meta-analysis of the N-G database. Furthermore, the itemwise tabulation explained twice as much of the variance as the subjectwise tabulation. However, the subjectwise analysis replicated the result commonly found with subjectwise analysis: Only recognition was significantly correlated with recognition success.

Why do subjectwise analysis and itemwise analysis yield different results? One explanation is that, in the subjectwise analysis, the correlation between $P(\text{RN}|\text{RC})$ and $P(\text{RC})$ is attenuated by a positive correlation between $P(\text{RN})$ and $P(\text{RC})$ [$r = .59$, $\text{CI } 95\% = (.32, .77)$], which does not exist for the itemwise tabulation [$r = -.03$, $\text{CI } 95\% = (-.31, .26)$]. Obviously, the more highly correlated $P(\text{RN})$ and $P(\text{RC})$ are, the less variance in $P(\text{RN}|\text{RC})$ will be accounted for by $P(\text{RC})$ independently of $P(\text{RN})$. One way of testing this explanation is to use Flexser's (1981) method of homogenizing 2×2 contingency tables to estimate subject and item covariance in the present data set. The former yielded a covariance estimate of .254, and the latter yielded a covariance estimate of .017. Therefore, according to Flexser, subject covariance inflated the dependency between $P(\text{RN})$ and $P(\text{RC})$, thus obscuring the relationship between $P(\text{RN}|\text{RC})$ and $P(\text{RC})$. At the same time, item covariance did not contribute to the dependency between $P(\text{RN}|\text{RC})$ and $P(\text{RN})$ and $P(\text{RC})$ in the itemwise tabulation (see also Metcalfe, 1991). So the subject covariance revealed by Flexser's method exaggerates the fit of the TW function to the data by obscuring the effect of $P(\text{RC})$, thus leaving only the correlation between $P(\text{RN}|\text{RC})$ and $P(\text{RN})$. This is exactly the kind of problem that Hintzman (1980) warned about.

Appendix B shows $P(\text{RN})$, $P(\text{RC})$, and $P(\text{RN}|\text{RC})$ for each item in the experiment. In addition, the probability that an item will be recalled but not recognized [i.e., $P(n\text{RN} \cap \text{RC})$], is included, because recognition failure rests on the occurrence of this joint event. Most surprisingly, 19 of the 47 items that are analyzed in the present experiment showed a $P(n\text{RN} \cap \text{RC})$ of .00. Therefore, the probability that each of these items will be recognized given that they are recalled is 1.0. In other words, these items do not contribute to recognition failure. Notice also that 10 other items had a negligible $P(n\text{RN} \cap \text{RC})$ of .03 (i.e., only 1 of 36 subjects recalled but failed to recognize the item). Averaging the probabilities of these 29 items, we found $P(\text{RN}) = .78$, $P(\text{RC}) = .28$, and $P(\text{RN}|\text{RC}) = .97$ (the TW function predicted .87). These items, therefore, contributed practically nothing to the recognition-failure phenomenon. The close agreement with the TW function indicated by Table 1 was therefore the sole responsibility of the use of the remaining 18 items.

What characterizes items that contribute strongly to the recognition-failure phenomenon? Look at the 14 items in Appendix A shown in bold. These are the word pairs in the list for which the A term is a noun and the B term is an adjective. In fact, 12 of the 18 items that exhibited non-negligible recognition failure are in this noun-adjective subset. To put it another way, the correlation between $P(\text{RN}|\text{RC})$ and $P(\text{RN})$ for these 14 noun-adjective items was $r = .83$, $\text{CI } 95\% = (.53, .94)$. For the remaining 33 items, the correlation was $r = .47$, $\text{CI } 95\% = (.15, .70)$.

The results for the noun-adjective pairs were not unexpected. Two earlier recognition-failure studies computed $P(\text{RN}|\text{RC})$ for the subset of items that were noun-adjective pairs. Bartling and Thompson (1977) computed

observed and predicted values of .64 and .73, respectively, for a deviation of $-.09$. Bartling (1992) computed observed and predicted values of .52 and .64, respectively, for a deviation of $-.12$. Here, the observed and predicted values were .75 and .80, respectively, for a deviation of $-.05$. So noun-adjective pairs consistently produce more recognition failure of recallable words than is predicted by the TW function. These results are undoubtedly related to the earlier finding that, in the recognition-failure paradigm, noun-adjective pairs produced both lower levels of recognition and higher levels of recall than for either verb-noun or noun-noun pairs (Olson, 1974). In fact, Olson found that the level of cued recall for the adjective was higher than the level of recognition, thus guaranteeing a nonzero level of recognition failure. DeVito (1975) found that the difference between cued recall and recognition for noun-adjective pairs was a function of the concreteness of the pair. By DeVito's criterion, all the pairs used here were concrete. So there was no reason to believe that associative strength between the noun and the adjective, or between members of other pairs, would predict the degree of recognition failure. In fact, associative strength did not predict recognition success [$r = -.08$, $CI\ 95\% = (-.36, .22)$]. This leaves the question of why concrete noun-adjective pairs produce higher levels of cued recall than recognition. An extensive discussion of this issue, which is beyond the scope of this report, may be found in studies examining the role of imagery in paired associate learning of noun-adjective pairs, including the study by Yuille, Paivio, and Lambert (1969). Briefly, it may be that high-imagery nouns are "conceptual pegs" from which related adjectives hang as "features." So the pre-existing relations between nouns and adjectives make the noun an effective cue for the noun-adjective pair.

So Hintzman's (1980, 1987, 1992) objection to the TW function is no longer completely theoretical. In fact, when some of the word pairs for which recognition failure was originally established are examined in detail, recognition failure is obtained for only a quarter of the pairs, which differ systematically from the three quarters for which recognition failure was not obtained. That is, recognition failure was obtained for noun-adjective pairs in which the adjective was a weak associate of the noun. The TW function was obtained when the different results for different types of word pairs were combined in a single subjectwise tabulation. Of course, this finding does not preclude the possibility that other types of items that are not noun-adjective pairs may also exhibit recognition failure. In fact, six pairs that were not noun-adjective pairs also exhibited nonnegligible recognition failure, and this was sufficient for the entire item set of 33 items to show a significant correlation for $P(RN|RC)$ with $P(RN)$ [$r = .49$, $CI\ 95\% = (.17, .71)$]. It demonstrates that preexisting relationships between items, rather than their contextual association in the study list, may provide a better explanation for how recognition failure occurs.

EXPERIMENT 2

The reexamination of the core data on which the TW function was originally based revealed that the function does not describe the data well enough to be considered a general law of memory. A more complete meta-analysis of the N-G database revealed that, in fact, recognition success is a function of both recognition and recall. This was also the case in the itemwise tabulation of the data in Experiment 1.

However, the most severe problem for the TW function was that most of the word pairs in Experiment 1 did not exhibit recognition failure. Instead, recognition failure was a characteristic of a subtype of word pairs (i.e., noun-adjective pairs). The TW function appeared to be an artifact of the ratio of the different subtypes of items in this particular item set. These findings support Hintzman's (1980) objection to contingency analysis in the recognition-failure paradigm. Furthermore, Rabinowitz, Mandler, and Barsalou (1977) and Bartling (1992) found that item-specific retrieval failure replicated across a series of experiments.

However, retrieval failure has been found with items other than noun-adjective pairs. Tulving and his colleagues might argue that the TW function has been replicated a great many times with many different kinds of materials and that it is incredible to imagine that all these experiments just happened to have the right mix of items that do and do not exhibit recognition failure in order to produce the TW function. In reply, we might point out that the impressive conformity of these many studies to the TW function may simply be the result of (1) a very large number of studies using familiar, weakly associated words as study materials, (2) a ubiquitous positive correlation between recognition and recall in the subjectwise tabulation, or (3) a lenient criterion for assessing the fit of the function to the data. Nevertheless, such a reply clearly falls in the realm of speculation. So more data are needed.

One problem with using words as study items in the recognition-failure paradigm is that they exhibit a host of preexisting associations that can be used to provide specific, noncontextual explanations of recognition failure. Therefore, in Experiment 2, pairs of familiar-foreign names were used because they had fewer preexisting associations.

If recognition failure is a function of the properties of specific items, then what properties of items are likely to contribute to recognition failure? Begg (1979) compared recognition failure for more recognizable but less recallable B terms with recognition failure for otherwise similar but less recognizable and more recallable B terms. He found more recognition failure for the less recognizable but more recallable B terms. We know that recallability is positively correlated with familiarity, but recognizability is negatively correlated with familiarity (e.g.,

Neely & Payne, 1983). Hence, increasing the familiarity of the B term should decrease $P(\text{RN})$ and increase $P(\text{RC})$, leading to a reduction in recognition failure. Conversely, decreasing the familiarity of the B term should increase $P(\text{RN})$ and decrease $P(\text{RC})$, leading to a reduction in recognition failure.

The items used in Experiment 2 were American-Norwegian and Norwegian-American name pairs, and the subjects were American college students. The names in each pair always shared the same first letter to increase the effectiveness of the A term as a cue for the B term. This was done to ensure a high enough level of recall so that the results could not be discounted because of poor performance in the cued recall task. It was assumed that the Norwegian names were less familiar to the subjects than were the American names. In the American-Norwegian condition, in which the less familiar Norwegian names served as the B terms, there should have been less recognition failure than in the Norwegian-American condition, in which the more familiar American names served as the B terms. Also, on the basis of the results of Neely and Payne (1983), it was expected that the American-Norwegian pairs would produce significantly less recognition failure than predicted by the TW function. When Neely and Payne used the first and last names of non-famous names (e.g., Martin CONWAY) in a recognition-failure task, they found that the observed level of recognition failure was less than predicted by the TW function. The American-Norwegian pairs (e.g., Mark MATS) were quite similar, consisting of a familiar American first name as the A term and a less familiar name as the B term, except that the B terms in Experiment 2 were less familiar than the B terms in Neely and Payne's study.

Another issue addressed by Experiment 2 was the number and kinds of name pairs for which recognition failure would occur. We suggested that recognition failure occurred in Experiment 1 for word pairs for which the A terms could be used to generate the B term through some preexisting semantic association. For American-Norwegian name pairs, such preexisting associations are less likely. But this does not rule out the possibility that the A term could be used to generate the B term on the basis of a perceptual similarity. To assess this possibility, both a subjective and an objective measure of the similarity between the terms were made. First, the subjects were asked to rate the similarity between the names. Second, we measured the overlap between the first four letters of the A term and the first four letters of the B term. Note that the amount of overlap had to be at least one letter because the names always shared the same first letter.

Method

Subjects and Materials. One hundred fifty subjects in an introductory course in psychology at Rutgers University participated in the experiment to satisfy a course requirement. These were English-speaking students between 18 and 30 years, who either were born in the U.S. or had lived in the U.S. since early childhood. None of the subjects had any knowledge of Norwegian or any other Scandinavian language.

The subjects were randomly assigned to one of two equally sized groups. In one group, the subjects studied Norwegian-American names. The study list for this group is presented in Appendix C. The other group studied the very same names with the members of a pair in reversed order (i.e., American-Norwegian names). The study list for both groups contained 60 pairs of names. The American names were all familiar to the student population at Rutgers. The length of any of these names did not exceed eight letters. The same restriction of length was imposed on the Norwegian names, none of which included a Scandinavian vowel missing in the English alphabet. The two names of a pair always shared the initial letter. The pairs of names were printed on two columns on a single page. A recognition test was constructed by randomly mixing target names with 60 new names of the same "nationality." The new names for the recognition test were selected according to the same principles that applied to the selection of target names. The 120 names in the recognition test were printed in three columns of 20 on two successive pages. The target names were randomly dispersed among the distractor names. The 60 names used as cues in the recall test were printed in two columns of 15 on two successive pages.

Procedure. To prevent floor effects in the cued recall test, the A members were presented in the same sequence as they had appeared in the study list, and the subjects were told to write next to each A word the corresponding B word. However, to avoid any primacy effects, we truncated the beginning 5 items before an analysis of the data (i.e., these items served as buffer items in the encoding episode). Recency effects were negligible since the subjects were given a retention interval of 1 week.

All subjects in both groups were given the following instructions:

You are participating in a memory experiment. Study the following pairs in preparation for a test which will be administered next week. You will eventually be asked to recognize and recall the second member of every pair. Also, study the first member of each pair since this name will help you to remember the second member. Please rate the following pairs for similarities (2 = very similar, 1 = similar, 0 = not at all similar.)

The subjects were given 15 min to perform the study task, and, during this episode, they were free to look back to previously studied pairs of names. After 1 week, they returned to be tested on recognition and recall. In the recognition test, the subjects were told to circle all the names they thought appeared in the study list. No time limit was imposed on either the recognition test or the recall test. However, the subjects were encouraged to finish each of the tests within 15 min. The subjects were tested individually or in small groups of 2 or 3 subjects at a time.

Results and Discussion

The data were tabulated by subjects and by items. Table 2 shows the mean and SD for each of the main variables for each group. Since we used a free version of the recognition test, d' was computed for each subject. This

Table 2
Retrieval Scores for the Norwegian-American and American-Norwegian Groups in Experiment 2

Variable	Norwegian-American			American-Norwegian		
	M	SD_{sw}	SD_{iw}	M	SD_{sw}	SD_{iw}
$P(\text{RN})$.63	.17	.12	.60	.19	.14
$P(\text{RC})$.34	.17	.24	.20	.16	.17
$P(\text{RN} \text{RC})_{\text{obs}}$.76	.18	.12	.90	.15	.11
$P(\text{RN} \text{RC})_{\text{TW}}$.75	.16	.10	.72	.17	.13
DevTW	.01	.11	.14	.18	.17	.16
d'	1.22	.78		1.37	.75	

measure, therefore, is included in the subjectwise descriptives of Table 2.

There was a negligible difference in the recognition score, and there was a nonsignificant difference in d' between the groups ($t = -1.57, p = .118$). In contrast, t tests for independent samples revealed significant differences between the groups, both itemwise and subjectwise, for $P(RC)$ and $P(RN|RC)$ ($p < .001$).

The data of the Norwegian-American group were clearly in agreement with the TW function, producing a negligible positive deviation of .01. In contrast, as predicted, the American-Norwegian group showed a remarkable positive deviation from this function. The DevTW of .18 for this group was clearly a significant deviation, with reference to both the critical ratio criterion used by Nilsson and Gardiner (1993) and the z used by Muter (1978) and Neely and Payne (1983), which, in this case, equaled 11.54. When the data were tabulated subjectwise, $P(RN)$ correlated with $P(RN|RC)$ for the Norwegian-American group [$r = .81, CI\ 95\% = (.71, .87)$] and for the American-Norwegian group [$r = .45, CI\ 95\% = (.24, .61)$]. $P(RC)$ correlated with $P(RN|RC)$ for the Norwegian-American group [$r = .44, CI\ 95\% = (.23, .60)$] and American-Norwegian group [$r = .13, CI\ 95\% = (-.10, .34)$].

The most remarkable correlations, however, were those between d' and the other variables, which are shown in Table 3. One might expect the correlations between d' and $P(RN)$, since $P(RN)$ is the hit rate, which is used in the computation of d' . However, the correlations between d' and $P(RC)$ were even higher.

Also, the predicted values according to Equations 1 and 2 served as independent variables in a regression analysis with the observed $P(RN|RC)$ as the dependent variable. For the Norwegian-American group, only Equation 2 entered the regression equation with a significant beta of .81, and the equation explained 66% of the variance. Also, for the American-Norwegian group, Equation 2 entered the equation, this time with a significant beta of .47, and the equation explained 22% of the variance.

When the data were tabulated itemwise, $P(RN)$ correlated with $P(RN|RC)$ for the Norwegian-American group [$r = .22, CI\ 95\% = (-.04, .45)$] and for the American-Norwegian group [$r = .09, CI\ 95\% = (-.18, .34)$].² $P(RC)$ correlated with $P(RN|RC)$ for the Norwegian-American group [$r = -.14, CI\ 95\% = (-.39, .13)$] and for the American-Norwegian group [$r = -.12, CI\ 95\% = (-.37, .15)$]. When the data from both groups were combined into a single data set, the correlation between $P(RN)$ and $P(RN|RC)$ was $r = .06, CI\ 95\% = (-.12, .24)$, and the

Table 3
Correlations Between d' and Other Variables in Experiment 2

Variable	Norwegian-American		American-Norwegian	
	d'	CI 95%	d'	CI 95%
$P(RN)$.71	(.58, .81)	.70	(.56, .80)
$P(RC)$.82	(.73, .88)	.81	(.71, .88)
$P(RN RC)_{obs}$.60	(.43, .73)	.28	(.06, .48)

Table 4
Correlations Between Rated Similarity Between Words of a Pair and Other Variables in Experiment 2

Variable	Norwegian-American		American-Norwegian	
	SIM	CI 95%	SIM	CI 95%
$P(RN)$.47	(.23, .65)	.41	(.16, .61)
$P(RC)$.81	(.69, .88)	.76	(.62, .85)
$P(RN RC)_{obs}$	-.10	(-.36, .17)	-.24	(-.47, .03)

correlation between $P(RC)$ and $P(RN|RC)$ was $r = -.29, CI\ 95\% = (-.45, -.10)$.

The itemwise analysis also included rated similarity (SIM; see Appendixes D and E) between the names of a pair as one of its variables. Table 4 shows the correlations between rated similarity (SIM) and $P(RN)$, $P(RC)$, and $P(RN|RC)$. Next, the independent variables $P(RN)$, $P(RC)$, and SIM were used to predict $P(RN|RC)$ for the Norwegian-American group and the American-Norwegian group. For the Norwegian-American group, all variables entered the regression equation with significant betas of .43 and $-.40$ for $P(RN)$ and $P(RC)$, respectively, and a nonsignificant beta of .08 for SIM. Equation 5, which summarizes the results for the Norwegian-American group, accounted for 15% of the variance. The analysis for the American-Norwegian group showed that only $P(RN)$ and SIM entered the regression equation, the former variable with a nonsignificant beta of .23 and the latter with a significant beta of $-.33$. Equation 6, which summarizes the results for this group when $P(RN)$ was forced in first, accounted for only 10% of the variance.

$$P(RN|RC) = .57 + .43P(RN) - .40P(RC) + .08SIM. \quad (5)$$

$$P(RN|RC) = .89 + .23P(RN) - .33SIM. \quad (6)$$

Also, for the itemwise tabulation of the data, we performed a stepwise regression analysis with the predicted values of Equations 1 and 2 as the independent variables. For the Norwegian-American group, only the predicted value of Equation 2 entered the equation, with a significant beta of .36, explaining 13% of the variance. For the American-Norwegian group, neither of the variables entered the regression equation.

We now turn to the issue of the number of name pairs for which recognition failure occurred and whether the amount of recognition failure was a function of the degree of name-pair similarity. The 12 name pairs in bold in Appendix C are the pairs for which the first four letters of the Norwegian name and the first four letters of the American name share at least three letters in common. We shall refer to these 12 pairs as the *high-similarity* subset and the remaining 43 pairs as the *low-similarity* subset. In the Norwegian-American condition, for the high-similarity subset, $P(nRN \cap RC) = .16$, and DevTW = $-.04$; for the low-similarity subset, $P(nRN \cap RC) = .06$ and DevTW = $.07$. The correlation between $P(RN|RC)$ and $P(RN)$ was $r = .84, CI\ 95\% = (.51, .95)$, for the high-similarity sub-

set, and it was $r = .18$, $CI 95\% = (-.12, .45)$, for the low-similarity subset. In the American-Norwegian condition, $P(nRN \cap RC) = .03$, and $DevTW = .13$, for the high-similarity subset, and $P(nRN \cap RC) = .01$, and $DevTW = .25$, for the low-similarity subset. The correlation between $P(RN \cap RC)$ and $P(RN)$ was $r = .54$, $CI 95\% = (-.05, .85)$, for the high-similarity subset, and it was $r = -.01$, $CI 95\% = (-.31, .29)$, for the low-similarity subset.

Again, in Experiment 2, recognition failure occurred for only a small number of items. The shape of the function that related $P(RN | RC)$ to $P(RN)$ in the subjectwise analysis was the result of collapsing across a large number of items that did not exhibit recognition failure and a small number that did.

GENERAL DISCUSSION

In the light of the results of the analysis presented here, let us consider again various claims made for the validity of the TW function, beginning with the strongest and proceeding systematically to the weakest. First, the TW function does not appear to be a universal law of memory. The majority of the word pairs in Experiment 1 and name pairs in Experiment 2 significantly deviated from the function. Even if the TW function is restricted to the subjectwise tabulations, it is not the best description of the N-G database for which it was originally conceived, and it fails to describe the results of the American-Norwegian condition of Experiment 2. Nor can the function be saved by restricting the range of recall and/or recognition values to which it is applied. Figure 2 clearly demonstrates that recall influences the TW function throughout the entire range of recall values.

Nor is it likely that the function can be saved by restricting the kinds of study materials to which it is applied. The results of both Experiment 1 and Experiment 2, for which different kinds of study materials were applied, indicate that as long as we select paired associates such that the B term cannot be generated from the A term, there will be a significant positive deviation from the TW function, regardless of the levels of recognition and recall.

Even if the specific TW function is abandoned and the more general claim that recognition success is solely a function of recognition; the same objections apply. So the function can not be saved merely by trying out other parameter values. The problem is that, contrary to the underlying assumption of the TW function, recognition failure is not the norm in the recognition-failure paradigm; rather, it is the exception. Generally, if the subject can recall a B term, then he/she can recognize that item as well. Notice that, in Experiment 2, the correlation between d' and recall was even higher than the correlation between recognition (i.e., the hit rate) and d' . So there is a strong contingent relationship between the recognition of a B term and its cued recall.

Of course, recognition failure undeniably does occur some of the time for some of the items. Let us consider how recognition failure occurs in the recognition-failure

paradigm. In fact, this paradigm is well named because it is peculiarly designed in order to produce recognition failure. To see why this is so, consider a paradigm in which subjects first studied A-B pairs, were then given a recognition test for the B words, and, finally, were given a recognition test for the A-B pairs. Obviously, we would expect recognition for the entire A-B pair to be superior to that for its part, the B term.³ Actually, this finding has been reported by Gardiner (1994). So, in the retrieval-failure paradigm, when a subject succeeds in generating the B term in response to the A term, the subject has succeeded in transforming the cued recall task into a recognition task for the entire A-B pair that is at least as easy as for the B term alone. So the amount of retrieval failure should be a measure of how much easier the A-B-recognition task is than the B-recognition task.

Remarkably, for most items there appears to be no difference in the discriminability of the B term versus the discriminability of the entire A-B pair in their respective retrieval tasks. The A-B pair appears to be more discriminable only when there exists between its members a relationship that exists independently of the pairing of the items in the study trial. In Experiment 1, this was a noun-adjective relationship; in Experiment 2, this was the number of letters the names had in common.

Originally, Tulving (1983) introduced the recognition-failure paradigm to demonstrate the validity of the encoding-specificity principle. He argued that recognition failure occurred because the items' contexts were the same during study and recall but were different during recognition. However, the results of the paradigm demonstrates the exact opposite. Consistent with Godden and Baddeley's (1975) well-replicated finding, the change in the context of the B term from isolation to a pairing with the A term usually has no effect on its recognizability. In fact, when recognition failure occurs, it is not because B has been studied in the presence of A; rather it is because there exists some relationship between A and B that exists independently of the fact that they have been seen together. So, when recognition failure occurs, it is not the result of a context-specific encoding but a context-independent association or relationship that subject can use to recognize the pair independently of whatever context-specific encoding occurred during the study trial.

Finally, the positive deviation of the American-Norwegian condition from the TW function points to what is possibly a critical design element in studies in which the TW function describes the results. This is the common use of item pairs that consist of familiar B terms and somewhat related or associated A terms. In fact, the fit of the TW function declines with the familiarity and relatedness of the study pairs. For example, Gardiner and Tulving (1980), Begg (1979), and Neely and Payne (1983) found that recognition failure for unrelated words was less than predicted by the TW function.

To restate these results another way, as the familiarity of the B term and its relatedness to the A term decrease, the amount of retrieval failure, and, hence, the probabil-

ity of a significant deviation from the TW function, increases. When the B term is sufficiently unfamiliar so that it is essentially a novel item, almost any encoding task sufficiently strong to produce cued recall for the item will produce recognition for the item as well, which is inconsistent with encoding specificity but is completely consistent with what is known about recognition and recall.

Other elements of the research process probably also contribute to the TW function. Good methodology involves the use of a sufficient number of items to establish generality over items. In practice, this has often meant the use of more than 20 items in a study list. However, when an effect is produced by only some of the items in a class, the use of a large number of items from the class all but guarantees that some of these items will be included.

It is therefore incumbent on the experimenter to demonstrate that pooled results generalize not only across subjects but across items as well, and that they are not the effect of a few specific items (Clark, 1973). Despite the warnings of Clark (1973), Hintzman (1980), and others, memory researchers have not taken care to do this. Other explanations have been offered for what have been interpreted as item-specific effects here. First, with equivalent study time, unfamiliar and unrelated items are less likely to be recalled than are more familiar and related items. As mentioned above, Gardiner and Nilsson (1990, 1993) have ruled out studies with low recall scores as encoding exceptions to the TW function. However, the results of the meta-analysis shown in Figure 2 demonstrate that the TW function is continuously related to recall throughout its full range. So there is no motivated reason for excluding studies on the basis of level of recall. Furthermore, significant deviations from the TW function have been found with recall ratios ranging from .20 (American-Norwegian condition of Experiment 2) to .23 (Begg, 1979) to .19-.29 (Gardiner & Tulving, 1980) to .33-.48 (Neely & Payne, 1983). If the TW function is descriptive only for recall levels above .20 or an even higher value, this begins to limit its generality.

There is another explanation for item-specific effects: that subjects use encoding strategies for unfamiliar or unrelated pairs different from the encoding strategies they use for familiar and related ones. Begg (1979) and Gardiner and Tulving (1980) suggested that retrieval failure occurs when subjects integrate the pair members during encoding. Thus, a lower level of retrieval failure for unfamiliar and unrelated items could be attributed to a lesser tendency for subjects to integrate them: According to this hypothesis, during the co-occurrence of the A and B terms on the study trial, the subject attempts to discover or construct an association that links the A and B terms. If such an association is created or found, then there will be some nonzero probability of generating the B term from the A term during the cued recall test. In fact, Gardiner and Tulving had some modest success in increasing retrieval failure by encouraging subjects to make linking sentences for otherwise unrelated pairs. However, if retrieval failure is determined by the tendency of subjects to en-

gage in item integration, then the TW function ultimately reflects the average tendency of subjects to engage in item integration for familiar weakly associated words under nonintegrative intentional learning conditions. Presumably, then, any change in the study task (e.g., a change in materials, instructions, or duration) that affects the tendency to engage in integration should affect the TW function. However, if this is the case, then the "exceptional conditions" are more extensive than the conditions under which the TW function may be observed. Furthermore, in order for the item-integration explanation to be different from the one advanced here, at a minimum, it must be shown that the linking association provided by integration must be noticed at encoding and that this linking association is not effective if first noticed at retrieval. Furthermore, the results of the classic study of Gardiner, Craik, and Birtwhistle (1972) demonstrate that this latter hypothesis is almost certainly false.

To summarize, when the restrictions on the kind of items for which the TW function has been found are taken into account, the conditions under which results consistent with the TW function has been observed are actually quite limited. Furthermore, Hintzman and Hartry (1990) have made criticisms of experiments conditionalizing recognition on fragment completion that complement the criticisms made here of experiments conditionalizing recognition on cued recall. So item analysis are applicable to studies of memory other than those into the role of context in recognition and recall.

REFERENCES

- ARIEF, M. (1996). Recognition failure: The influence of semantic cue-target integration—a short note. *European Journal of Cognitive Psychology*, *8*, 205-214.
- BARTLING, C. A. (1992). List-subset effects and the Tulving-Wiseman function. *Bulletin of the Psychonomic Society*, *30*, 131-134.
- BARTLING, C. A., & THOMPSON, C. P. (1977). Encoding specificity: Retrieval asymmetry in the recognition failure paradigm. *Journal of Experimental Psychology: Human Learning & Memory*, *3*, 690-700.
- BEGG, I. (1979). Trace loss and the recognition failure of unrecalled words. *Memory & Cognition*, *7*, 113-123.
- CLARK, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, *12*, 335-359.
- COHEN, R. L. (1985). On the generality of the laws in memory. In L.-G. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 247-277). Hillsdale, NJ: Erlbaum.
- DEVITO, C. (1975). Encoding specificity and integration of word pairs. *Bulletin of the Psychonomic Society*, *5*, 215-216.
- FLEISHER, A. J. (1981). Homogenizing the 2 × 2 contingency table: A method for removing dependencies due to subject and item differences. *Psychological Review*, *88*, 327-339.
- FLEISHER, A. J., & TULVING, E. (1978). Retrieval independence in recognition and recall. *Psychological Review*, *85*, 153-171.
- GARDINER, J. M. (1994). The Tulving-Wiseman law and recognition failure of recognizable words. *European Journal of Cognitive Psychology*, *6*, 93-105.
- GARDINER, J. M., CRAIK, F. I. M., & BIRTWHISTLE, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning & Verbal Behavior*, *11*, 778-783.
- GARDINER, J. M., & NILSSON, L.-G. (1990). *Relation between recognition and recall: The Tulving-Wiseman law* (Umeå Psychological Rep. No. 205). Umeå, Sweden: University of Umeå.

- GARDINER, J. M., & NILSSON, L.-G. (1993). Mathematical constraints and the Tulving-Wiseman law: A rejoinder. *Memory*, 1, 219-229.
- GARDINER, J. M., & TULVING, E. (1980). Exception to recognition failure of recallable words. *Journal of Verbal Learning & Verbal Behavior*, 19, 194-209.
- GODDEN, D. R., & BADDELEY, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, 66, 325-332.
- HAYMAN, C. A. G., & TULVING, E. (1989). Contingent dissociation between recognition and fragment completion: The method of triangulation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 15, 228-240.
- HINTZMAN, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, 87, 398-410.
- HINTZMAN, D. L. (1987). Recognition and recall in MINERVA 2: Analysis of the "recognition-failure" paradigm. In P. E. Morris (Ed.), *Modelling cognition* (pp. 215-229). New York: Wiley.
- HINTZMAN, D. L. (1992). Mathematical constraints and the Tulving-Wiseman law. *Psychological Review*, 99, 536-542.
- HINTZMAN, D. L., & HARTRY, A. L. (1990). Item effects in recognition and fragment completion: Contingency relations vary for different subsets of words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 16, 955-969.
- JONES, G. V. (1984). Analyzing recognition and recall. *Behavioral & Brain Sciences*, 7, 242-243.
- METCALFE, J. (1991). Recognition failure and the composite memory trace in CHARM. *Psychological Review*, 98, 529-553.
- MUTER, P. (1978). Recognition failure of recallable words in semantic memory. *Memory & Cognition*, 6, 9-12.
- NEFLY, J. H., & PAYNE, D. G. (1983). A direct comparison of recognition failure rates for recallable names in episodic and semantic memory tests. *Memory & Cognition*, 11, 161-171.
- NILSSON, L.-G., & GARDINER, J. M. (1993). Identifying exceptions in a database of recognition failure studies from 1973 to 1992. *Memory & Cognition*, 21, 397-410.
- NILSSON, L.-G., LAW, J., & TULVING, E. (1988). Recognition failure of recallable unique names: Evidence for an empirical law of memory and learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 14, 266-277.
- OLSON, A. M. (1974). The differential effects of syntactical pairings on cued recall and recognition. *Bulletin of the Psychonomic Society*, 3, 232-233.
- RABINOWITZ, J. C., MANDLER, G., & BARSALOU, L. W. (1977). Recognition failure: Another case of retrieval failure. *Journal of Verbal Learning & Verbal Behavior*, 16, 639-663.
- RIEFER, D. M., & BATCHELDER, W. H. (1995). A multinomial modeling analysis of the recognition-failure paradigm. *Memory & Cognition*, 23, 611-630.
- TULVING, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.
- TULVING, E., & THOMSON, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352-373.
- TULVING, E., & WISEMAN, S. (1975). Relation between recognition and recognition failure of recallable words. *Bulletin of the Psychonomic Society*, 6, 79-82.
- WATKINS, M. J., & TULVING, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General*, 104, 5-29.
- YUILLE, J. C., PAIVIO, A., & LAMBERT, W. E. (1969). Noun and adjective imagery and order in paired associate learning by French and English subjects. *Canadian Journal of Psychology*, 23, 459-466.

NOTES

1. Equation 2, due to an intersection of .28 with the y-axis, may have solutions above 1.0. However, since we only consider data points within the unit graph space, predicted scores that exceeded 1.0 were set equal to 1.0.

2. Why were the itemwise correlations between $P(RN)$ and $P(RN:RC)$ so much lower in Experiment 2 than in Experiment 1? One possibility is that many more items were hardly ever recalled in Experiment 2 than in Experiment 1. In Experiment 1, 4 of 48 items were recalled by less than 14% of the subjects (see Appendix B); however, in the Norwegian-American condition of Experiment 2, 10 of 55 items were recalled by less than 14% of the subjects (see Appendix E). At low levels of $P(RC)$, $P(RN:RC)$ is necessarily based on very few observations and, hence, is mostly error variance, thus reducing the reliability of the correlations. When items recalled by less than 14% of the subjects were excluded, the magnitude and reliability of the correlations were similar to the levels in Experiment 1.

3. Equally obviously, we could design materials in which the A terms would camouflage otherwise distinct B terms, and, thus, recognition of the whole would be less than for its part. However, this is obviously not the case for the materials that have been ubiquitously used in the recognition-failure paradigm.

APPENDIX A
English and Norwegian Word Pairs
Presented in the Study List of Experiment 1

Item No.	English	Norwegian	Association Rating	Distractors
1	plant-BUG	plante FLUE	1.72	WASP WORM
2	wish-WASH	søl-VASKE	2.56	RINSE DRY
3	hope-HIGH	håpe-HØY	1.08	SKY DESIRE
4	stem-SHORT	stamme-KORT	1.12	SMALL HEAVY
5	whisky-WATER	whisky-VANN	2.36	RIVER STREAM
6	moth-FOOD	munn-MAT	2.90	MEAL SOUP
7	cabbage-ROUND	kål-RUND	2.48	SHAPE RECTANGLE
8	glass-HARD	glass-HARD	2.08	SILENT SOFT
9	country-OPEN	land-ÅPEN	1.52	LOCKED CLOSED
10	tool-HAND	redskap-HÅND	2.26	NAIL TOUCH
11	memory-SLOW	hukommelse-TREG	2.14	QUICK START
12	covering-COAT	dekke-FRAKK	1.66	FUR CLOAK
13	barn-DIRTY	låve-SKITTEN	1.42	SHINE POLISH
14	spider-BIRD	edderkopp-FUGL	1.38	SIGN FISH
15	crust-CAKE	skorpe-KAKE	2.00	BREAD BISCUIT
16	deep-SLEEP	dyp-SØVN	2.58	PILLOW HAMMOCK
17	train-BLACK	tog-SORT	1.32	CLOUT NIGHT
18	mountain-TREE	fjell-TRE	1.78	FALL LEAF
19	cottage-LOVE	hytte-FRIHET	2.14	DEMAND PHONE
20	art-GIRL	kunst-JENTE	1.42	FRIEND STUDENT
21	adult-WORK	voksen-ARBEIDE	2.46	LEADER BIRTH
22	brave-WEAK	modig-SVAK	1.56	IMPORTANT POWERFUL
23	door-RED	port-RØD	1.52	ORANGE VIOLET
24	roll-RUG	rull-TEPPE	2.32	RAG CURTAIN
25	think-STUPID	tenke-DUM	1.86	QUIET DEAF
26	exist-BEING	eksistere-MENNESKE	2.62	PERSON CREATURE
27	home-SWEET	hjem-BRA	1.98	GOOD NICE
28	grasp-BABY	gripe-BABY	1.84	KID BROTHER
29	butter-SMOOTH	smør-GLATT	2.16	STRONG CAUTIOUS
30	drink-SMOKE	drikk-ROYK	2.04	TASTE SMELL
31	beat-PAINE	slå-SMERTE	2.94	FEELING EVIL
32	cloth-SHEEP	klær-SAU	2.22	GOAT DEER
33	swift-GO	rask-GÅ	2.00	STOP BEGIN
34	lady-QUEEN	kvinne-DRONNING	2.70	PRINCE MINISTER
35	blade-CUT	egg-SKJÆRE	1.50	STAB SPLIT
36	ground-COLD	jord-KALD	1.84	WARM CHILLY
37	head-LIGHT	hode-LYS	1.58	DARKNESS DIMNESS
38	bath-NEED	bade-BEHOV	1.58	HELP REQUIREMENT
39	cheese-GREEN	ost-GRØNN	1.60	YELLOW BROWN
40	stomach-LARGE	mage-STOR	2.26	WIDE ENORMOUS
41	sun-DAY	søl-DAG	2.58	WEEK MONTH
42	pretty-BLUE	pen-BLÅ	1.36	GREY PURPLE
43	cave-WET	grotte-VÅT	2.14	FOG MIST
44	whistle-BALL	fløyte-BALL	1.82	RACKET NET
45	noise-WIND	støy-VIND	1.88	STORM TORNADO
46	glue-CHAIR	klister-STOL	1.02	BENCH STOOL
47	command-MAN	kommando-MANN	2.02	BOY GUY
48	fruit-FLOWER	frukt-BLOMST	2.34	BLOSSOM GRAIN

Note—Word pairs which are noun-adjective pairs in Norwegian are printed in bold. Distractors are new words that were presented together with B (the target word) in the recognition test.

APPENDIX B
Retrieval Probabilities for Items in Experiment I

Item No.	$P(RN)$	$P(RC)$	$P(nRN \cap RC)$	$P(RN RC)_{obs}$	$P(RN RC)_{TW}$	DevTW
1	.83	.11	.00	1.00	.90	.10
2	.58	.22	.00	1.00	.70	.30
3	.67	.11	.00	1.00	.78	.22
4	.75	.36	.11	.69	.85	-.16
5	.67	.92	.28	.70	.79	-.09
6	.64	.33	.11	.67	.75	-.09
7	.67	.72	.19	.73	.78	-.05
8	.50	.72	.36	.50	.63	-.13
9	.67	.25	.03	.89	.78	.11
10	.81	.39	.03	.92	.88	.04
11	.75	.31	.06	.80	.85	-.05
12	.67	.33	.00	1.00	.78	.22
13	.86	.42	.06	.87	.92	-.05
14	.61	.44	.14	.63	.73	-.10
15	.56	.36	.08	.79	.69	.10
16	.78	.00	.00	-	.87	-
17	.72	.58	.14	.76	.82	-.06
18	.81	.33	.00	1.00	.89	.11
19	.82	.31	.00	1.00	.90	.10
20	.97	.39	.00	1.00	.99	.01
21	.72	.28	.00	1.00	.82	.18
22	.75	.19	.03	.86	.85	.01
23	.83	.50	.03	.94	.90	.04
24	.61	.50	.08	.83	.73	.10
25	.86	.17	.00	1.00	.92	.08
26	.86	.42	.00	1.00	.92	.08
27	.39	.17	.08	.50	.51	-.01
28	.69	.28	.03	.90	.80	.10
29	.56	.42	.14	.67	.69	-.02
30	.69	.11	.00	1.00	.80	.20
31	.69	.22	.00	1.00	.80	.20
32	.81	.19	.00	1.00	.89	.11
33	.67	.14	.00	1.00	.78	.22
34	1.00	.61	.00	1.00	1.00	.00
35	.78	.19	.00	1.00	.86	.14
36	.81	.39	.06	.86	.89	-.03
37	.89	.19	.00	1.00	.94	.06
38	.61	.17	.00	1.00	.73	.27
39	.81	.42	.08	.80	.89	-.09
40	.69	.44	.17	.63	.80	-.18
41	.81	.31	.03	.91	.89	.02
42	.81	.28	.03	.90	.89	.01
43	.81	.53	.08	.84	.89	-.05
44	.81	.25	.03	.89	.89	.00
45	.81	.14	.00	1.00	.89	.11
46	.67	.22	.03	.88	.78	.10
47	.72	.58	.08	.86	.82	.03
48	.75	.28	.03	.90	.84	.06

Note—Item numbers are consistent with item numbers in Appendix A.

APPENDIX C
Norwegian-American Name Pairs
Presented in the Study List of Experiment 2

Item No.	Name Pair	Item No.	Name Pair
1	Oda-Oprah	31	Gaute-Gary
2	Petra-Patty	32	Hedvik-Helen
3	Jostein-Justin	33	Espen-Eva
4	Asle-Alfred	34	Aina-Ann
5	Tordis-Toby	35	Trine-Tina
6	Frode-Fred	36	Laila-Lauren
7	Jesper-Jessica	37	Kaja-Kathleen
8	Agnar-Arthur	38	Anders-Alex
9	Vivi-Vivian	39	Bente-Benjamin
10	Mette-Michael	40	Stian-Stan
11	Dag-Doug	41	Gunnar-Glen
12	Dorte-Dorothy	42	Einar-Ellen
13	Gudveig-George	43	Mons-Martha
14	Arne-Arnold	44	Ingvil-Irene
15	Marit-Marie	45	Gunval-Gwen
16	Arnt-Arlene	46	Synne-Sarah
17	Kjetil-Keith	47	Bernt-Bob
18	Maren-Mary	48	Guro-Gregory
19	Dagfinn-Donald	49	Mats-Mark
20	Signe-Susan	50	Enok-Elaine
21	Oddvar-Oscar	51	Magnus-Matthew
22	Jorunn-Justine	52	Tonje-Thomas
23	Vidar-Victor	53	Stine-Stacy
24	Venche-Victoria	54	Pelle-Pamela
25	Line-Lynne	55	Erling-Edward
26	Solveig-Sheila	56	Unni-Ursula
27	Jens-John	57	Turid-Ted
28	Lene-Lenny	58	Aksel-Andrew
29	Runar-Raymond	59	Anja-Angela
30	Stig-Steven	60	Trude-Tamara

Note—The pairs in bold are those in which the A and B terms share at least three of the first four letters.

APPENDIX D
Retrieval Probabilities for the Norwegian-American Group in Experiment 2

Item No.	SIM	$P(RN)$	$P(RC)$	$P(nRN \cap RC)$	$P(RN RC)_{obs}$	$P(RN, RC)_{TW}$	DevTW
6	2.0	.72	.95	.28	.71	.82	-.12
7	1.2	.65	.34	.08	.77	.77	.00
8	0.9	.68	.25	.04	.84	.79	.05
9	2.4	.91	.95	.09	.90	.95	-.04
10	0.6	.63	.17	.04	.77	.75	-.02
11	1.8	.80	.88	.14	.84	.88	-.05
12	2.0	.62	.86	.33	.62	.74	-.12
13	0.6	.63	.33	.11	.68	.75	-.07
14	2.0	.65	.63	.20	.69	.77	-.08
15	1.8	.68	.53	.11	.80	.79	.01
16	1.0	.63	.17	.04	.77	.75	.02
17	0.7	.67	.55	.13	.76	.78	-.02
18	1.6	.74	.46	.07	.86	.83	.02
19	0.6	.45	.21	.04	.81	.57	.24
20	0.6	.47	.14	.01	.91	.60	.31
21	1.3	.80	.62	.09	.85	.88	-.03
22	0.8	.84	.21	.03	.88	.91	-.03
23	1.4	.70	.63	.14	.77	.80	-.03
24	0.6	.63	.33	.07	.80	.75	.05
25	2.2	.71	.74	.17	.77	.81	-.05
26	0.6	.47	.07	.01	.80	.60	.20
27	1.0	.57	.40	.14	.63	.69	-.06
28	2.3	.63	.70	.25	.64	.73	-.08
29	0.6	.59	.20	.08	.60	.71	-.11
30	0.9	.61	.18	.07	.64	.73	-.08
31	0.9	.55	.25	.08	.68	.68	.00
32	0.8	.50	.32	.17	.46	.63	-.17
33	0.6	.80	.18	.04	.79	.88	-.10
34	1.8	.68	.34	.05	.85	.79	.05
35	1.6	.70	.50	.13	.74	.80	-.07
36	1.0	.58	.25	.00	1.00	.70	.30
37	0.7	.53	.30	.05	.83	.65	.17
38	0.6	.55	.07	.00	1.00	.68	.32
39	1.3	.71	.59	.09	.84	.81	.03
40	2.2	.61	.66	.21	.68	.73	-.05
41	0.7	.61	.20	.03	.87	.73	.14
42	0.8	.58	.18	.03	.86	.70	.16
43	0.6	.42	.12	.03	.78	.54	.24
44	0.6	.57	.18	.07	.64	.69	-.05
45	0.8	.67	.18	.03	.86	.78	.07
46	0.7	.55	.18	.03	.86	.68	.18
47	0.6	.36	.12	.01	.89	.47	.42
48	0.9	.71	.07	.01	.80	.81	-.01
49	0.9	.49	.21	.05	.75	.61	.14
50	0.6	.68	.11	.03	.75	.79	-.04
51	0.9	.65	.12	.00	1.00	.77	.23
52	1.0	.49	.11	.01	.88	.61	.26
53	1.0	.46	.20	.05	.73	.59	.15
54	1.1	.54	.20	.07	.67	.66	.00
55	0.7	.37	.13	.07	.50	.48	.02
56	0.8	.65	.59	.13	.78	.77	.01
57	1.1	.47	.12	.03	.78	.60	.18
58	0.7	.65	.14	.05	.64	.77	-.13
59	2.2	.74	.46	.11	.77	.84	-.07
60	0.6	.68	.18	.00	1.00	.79	.21

Note—SIM = rated similarity. Item numbers are consistent with the item numbers in Appendix C. The beginning five items (Items 1-5) were truncated; therefore, the analysis was undertaken for Items 6-60.

APPENDIX E
Retrieval Probabilities for the American-Norwegian Group in Experiment 2

Item No.	SIM	$P(\text{RN})$	$P(\text{RC})$	$P(\text{nRN} \cap \text{RC})$	$P(\text{RN} \text{RC})_{\text{obs}}$	$P(\text{RN} \text{RC})_{\text{pred}}$	DevTW
6	1.9	.76	.48	.04	.92	.77	.15
7	1.4	.80	.35	.03	.93	.88	.05
8	1.0	.79	.00	.00	1.00	.88	.12
9	2.3	.91	.85	.04	.96	.95	.00
10	0.7	.76	.20	.00	1.00	.85	.15
11	2.0	.73	.65	.06	.90	.83	.07
12	2.1	.71	.56	.07	.87	.82	.05
13	0.7	.70	.14	.00	1.00	.81	.19
14	1.9	.84	.48	.02	.95	.91	.04
15	1.9	.66	.24	.02	.89	.78	.11
16	1.3	.65	.21	.04	.82	.76	.06
17	0.7	.60	.16	.00	1.00	.72	.28
18	1.8	.44	.16	.06	.62	.56	.05
19	0.6	.66	.13	.02	.90	.78	.12
20	0.7	.66	.09	.00	1.00	.78	.22
21	1.9	.73	.30	.02	.92	.83	.09
22	0.7	.34	.04	.00	1.00	.45	.55
23	1.5	.70	.14	.02	.91	.81	.10
24	0.6	.59	.13	.02	.80	.71	.09
25	2.1	.56	.53	.13	.76	.69	.07
26	0.6	.65	.05	.00	1.00	.76	.24
27	1.2	.65	.21	.01	.94	.76	.18
28	2.3	.65	.31	.05	.84	.76	.08
29	0.8	.39	.13	.02	.80	.46	.34
30	1.0	.54	.14	.00	1.00	.67	.33
31	1.2	.53	.09	.00	1.00	.66	.34
32	0.8	.54	.10	.03	.63	.67	-.04
33	0.7	.55	.10	.01	.88	.68	.20
34	1.9	.59	.23	.00	1.00	.71	.29
35	1.5	.75	.35	.00	1.00	.84	.16
36	0.9	.57	.19	.00	1.00	.70	.30
37	0.8	.59	.24	.02	.89	.71	.18
38	1.0	.57	.21	.01	.94	.70	.24
39	1.3	.53	.21	.01	.94	.66	.28
40	2.1	.44	.19	.03	.87	.56	.30
41	0.9	.69	.11	.01	.89	.80	.09
42	1.0	.39	.05	.00	1.00	.46	.54
43	0.7	.49	.04	.00	1.00	.61	.39
44	0.7	.64	.04	.00	1.00	.75	.25
45	0.8	.39	.04	.00	1.00	.46	.54
46	0.9	.43	.08	.00	1.00	.55	.45
47	0.8	.48	.08	.02	.67	.60	.07
48	0.9	.35	.06	.00	1.00	.46	.54
49	1.1	.63	.14	.00	1.00	.74	.26
50	0.6	.60	.05	.00	1.00	.72	.28
51	1.1	.57	.06	.00	1.00	.70	.30
52	1.0	.59	.09	.00	1.00	.71	.29
53	1.2	.56	.06	.00	1.00	.69	.31
54	1.1	.53	.14	.03	.82	.65	.17
55	1.0	.35	.05	.00	1.00	.46	.54
56	0.9	.53	.24	.01	.95	.65	.30
57	1.2	.48	.06	.02	.60	.60	.00
58	0.7	.33	.05	.00	1.00	.43	.57
59	2.0	.76	.30	.03	.88	.85	.02
60	0.7	.51	.09	.00	1.00	.64	.36

Note—SIM = rated similarity. Item numbers are consistent with item numbers in Appendix C, but the order of names is reversed. The beginning five items (Items 1-5) were truncated; therefore, the analysis was undertaken for Items 6-60.

(Manuscript received December 4, 1996;
revision accepted for publication May 8, 1997.)