

## Retrieval independence in successive recognition tasks

Arnold L. Glass

*Rutgers University, New Brunswick, New Jersey, USA*

Arild Lian and Tore Helstrup

*University of Oslo, Oslo, Norway*

In three experiments participants studied AB word pairs and completed two recognition tests. In the first recognition test, which was included in all three experiments, the B word had to be discriminated from two distractors that did not appear on the study list. In Experiment 1, in the second recognition test, an AB target was compared with distractors composed of words not on the study list. In Experiment 2, in the second recognition test, an AB target had to be discriminated from two other pairs that were created by randomly re-pairing A and B words that appeared on the study list. In Experiment 3, on the second recognition test, words from the study list were systematically re-paired to form distractors that contained either the same A term or the same B term as the target pair. Recognition of the B word on the first test was always at least partly independent of recognition of the AB pair on the second test. Even when recognition judgements were restricted to those for which the participants were most confident, all experiments demonstrated significant retrieval independence between the two tests.

It seems reasonable to assume that there should be a high degree of dependence between the items retrieved when successive retrieval tests of the same study list are made. However, this is not always the case. When a paired associate (AB) learning task has been followed first by a test of recognition of the B word and then by a test of recall of B given A as a cue there is considerable independence between recognition and recall of the B word (Tulving, 1983). Among the various attempts to explain the relative independence between recognition and cued recall we find the retrieval independence model of Flexser and Tulving (1978). This is a mathematical model with two key assumptions. The first is what they call *the trace identity assumption*. This is that there is a single pool of features that may be activated at encoding or retrieval. The second is the *retrieval independence assumption*. An independent set of features is sampled from the pool during encoding and during each of the retrieval tests. Notice that the higher the proportion of the feature pool sampled during each encoding and retrieval

---

Requests for reprints should be sent to Arnold L. Glass, Psychology Department, Rutgers University, New Brunswick NJ 8903, USA. Email: [aglass@rci.rutgers.edu](mailto:aglass@rci.rutgers.edu)

event the greater the overlap in features sampled and hence the greater the dependence between the results of successive retrieval tests. Conversely, the lower the proportion of the feature pool sampled during each encoding and retrieval event the lower the overlap in features sampled and hence the greater the independence between the results of successive retrieval tests. Flexser and Tulving selected parameter values for sample sizes that generated predictions that conformed to the observed dependency between recognition and cued recall.

Flexser and Tulving (1978) showed that the key assumption that made it possible for their model to fit the data was the retrieval independence assumption. They showed that assuming even 25% overlap in the features sampled on successive retrieval tests predicted greater dependence between the results of the tests than that observed. Flexser and Tulving did not say what kinds of cues would activate dependent feature sets and hence would produce more dependence than that observed when B recognition was followed by cued recall in response to A. However, Gardiner (1994) suggested that if a B recognition test were followed by an AB recognition test then these cues would sample overlapping feature sets because they both contained B as a cue and were both recognition tests. According to this *informational overlap hypothesis*, there should be more dependence between B recognition and AB recognition than between B recognition and AB recall in response to A. However, Gardiner found the same degree of independence between the results of the B and AB recognition tests as that previously observed between B recognition and cued recall in response to A. To account for this finding Gardiner proposed that a second retrieval test would show the same degree of independence from the first as that typically observed when B recognition is followed by cued recall in response to A, as long as the second retrieval test contains some information not available in the first test. In his experiments this information would be the A word, which is only present in the second recognition test. Because he called the additional information *contextual information* Gardiner called his explanation the *contextual account* of retrieval independence between successive tests. Gardiner did not describe the contextual account in any more detail than that presented here. So the explanation for the degree of independence observed between the results of successive retrieval tests is vague.

The purpose of this report is to examine key assumptions underlying extant explanations of independence in successive recognition tests. To begin, we shall consider in more detail why Gardiner's (1994) finding that there is no more independence between successive recognition tests than between a recognition test followed by a cued recall test challenges the retrieval independence assumption. Second, we apply a quantitative model of the encoding and retrieval processes involved in recognition to Gardiner's results and see that the contrary *retrieval consistency* assumption provides an adequate account of his results. Third, we consider the implications of the remember/know distinction for comparing successive recognition tests with recognition followed by cued recall. The remember/know distinction suggests the possibility that response criterion differences muddied the comparison across tasks. Finally, we report the results of three experiments that directly test the retrieval independence and trace identity assumptions. We make use of the quantitative model introduced in conjunction with Gardiner's data to evaluate the results.

## Testing retrieval independence: Successive recognition versus recognition followed by recall

Gardiner (1994) reported three experiments in which a list of AB study pairs was followed by two recognition tests. A recognition test for the B item was followed by a second recognition test for the entire AB pair. In the first two experiments he used the original English-language word pairs from Tulving and Thomson (1973). The experiments involved two alternative study lists, each with 24 word pairs. There were two groups of participants, one for each study list. The first recognition test was created by combining the B words from both lists. The second test involved presentation of the B words along with their corresponding A words mixed together with new word pairs consisting of words that had not been presented in the study task. In Experiment 1 Gardiner varied the retention interval, and in Experiment 2 he varied the rate of presentation of AB items in the study task.

As mentioned above, Gardiner's (1994) experiments were performed in the context of previous research in which a recognition test for the B word was followed by a recall test in which the A word was presented as a cue for the B word. Consider the relationship between the probability of recognition of the B word,  $P(\text{RnB})$ , and the probability of recall of the B word given the A word,  $P(\text{RcB})$ . In a large number of studies reviewed by Nilsson and Gardiner (1993),  $P(\text{RnB} | \text{RcB})$  had a value intermediate between pure independence and dependence. The value of  $P(\text{RnB} | \text{RcB})$  was less than 1.0, thus indicating a degree of independence between recognition and cued recall of the B word.

One explanation for the degree of independence between  $P(\text{RnB})$  and  $P(\text{RcB})$  is that both tests make use of different retrieval cues, B in the first test and A in the second, and these have independent probabilities of activating the representation in memory. If retrieval independence is the result of nonoverlap between the cues on successive retrieval tasks then increasing the similarity between the retrieval cues on the two tasks should increase the degree of dependence between the retrieval tasks. Since the successive recognition tasks, in which the B word alone and then the entire AB pair are recognized, share the B word there should be greater dependence between the results of the successive B/AB recognition tasks than for recognition of B followed by cued recall of B by A, which do not share any cue. That is, the probability of recognizing the B word given that the entire AB pair was recognized,  $P(\text{RnB} | \text{RnAB})$ , should be significantly greater than the value of  $P(\text{RnB} | \text{RcB})$  obtained in most experiments of recognition failure of recallable words (Nilsson & Gardiner, 1993; Tulving & Wiseman, 1975), where the retrieval tasks do not share a cue.

In fact, Gardiner (1994) found the same degree of independence between B and AB recognition as that previously observed between B recognition and B recall as the response to the A cue in traditional experiments on recognition failure of recallable words. The results of his experiments are shown in Table 1. Rows 1 and 2 show the results of the two conditions of his first experiment, which differed only in a retention interval between the study task and first recognition test of 10 min versus 4 days. Rows 3 and 4 show the results of the two conditions of his second experiment, which differed only in the study time for each study item of 1 s versus 5 s. Table 1 also shows the values of Yule's  $Q$ , which, as described below, provides a measure of

TABLE 1  
Observed and predicted results of Gardiner's (1994) study

Exp.	Condition	Obs./ Pre.	Yule's Q	Probability			Parameters									
				RnB	RnAB	RnB   RnAB	E	S <sub>1</sub>	S <sub>3</sub>	S <sub>2</sub>	S <sub>1</sub>					
1	10 min	Obs.	.69	.45	.77	.53	.72	1.00	.63	.63	.00					
		Pre.	.62	.45	.77	.53										
	4 days	Obs.	.65	.32	.62	.51										
	4 days	Pre.	.64	.32	.62	.5						.55	.92	.70	.70	.00
2	1 s	Obs.	.51	.49	.52	.62	.49	.80	.78	.46	.00					
		Pre.	.60	.49	.52	.62										
	5 s	Obs.	—	.73	.88	.83										
	5 s	Pre.	.87	.73	.87	.81						.73	1.00	1.00	.85	.00

Obs. = observed; Pre. = predicted.

independence between  $P(RnB)$  and  $P(RnAB)$  that can be used for comparisons across conditions and experiments. As described below, both a Yule's  $Q$  of  $-1.0$  and that of  $+1.0$  mean complete dependence, while a Yule's  $Q$  of  $0$  means complete independence between the tests compared. Unfortunately, if there are no observations in one of the cells in the two by two contingency table on which Yule's  $Q$  is based, the substantial interpretation of Yule's  $Q$  is unclear (Bishop, Fienberg, & Holland, 1975). For Gardiner's 5-s presentation rate in his second experiment the  $P(RnB \cap RnAB)$  cell (i.e., the trials on which  $B$  was recognized but  $AB$  was not) was empty, and hence Yule's  $Q$  was not computed for this condition.

## Modelling independence in successive recognition tasks

Let us consider how the learning and recognition processes might be modelled and the results of Gardiner's (1994) first two experiments predicted. We begin with a model that is consistent with some of the assumptions of Flexser and Tulving (1978), Gardiner (1994) and, as we shall see, other theorists. Consider what happens when an  $AB$  word pair is presented on a study trial. As shown in Figure 1(a), there is some probability,  $E$ , that  $A$  will be encoded, and the same probability,  $E$ , that  $B$  will be encoded. Hence, the probability that both  $A$  and  $B$  are encoded is  $E^2$ . If both  $A$  and  $B$  are encoded then there is some probability,  $F$ , that an association between  $A$  and  $B$  will be formed and hence the pair  $AB$  will be encoded. We use  $(AB)$  to indicate that both terms of the pair have been encoded in a single representation. This means that the participant will remember seeing the entire pair as a unit. For example, if  $A_1B_1, A_2B_2, A_3B_3$  are presented on the first three trials then encoding the association between  $A_1$  and  $B_1$  makes it possible for someone to discriminate the target  $A_1B_1$  from the distractors  $A_2B_3$  and  $A_3B_2$  on a subsequent forced-choice recognition test. We may contrast the encoding of both terms in a single representation with the possibility that each term is encoded in a separate representation, which may occur when  $F < 1.0$ . We use  $(A,B)$  to indicate that each term of the pair has been encoded in a separate representation. In this case a participant would remember seeing the  $A$  term and remember seeing the  $B$  term, but would not remember seeing the pair together.

## Representation

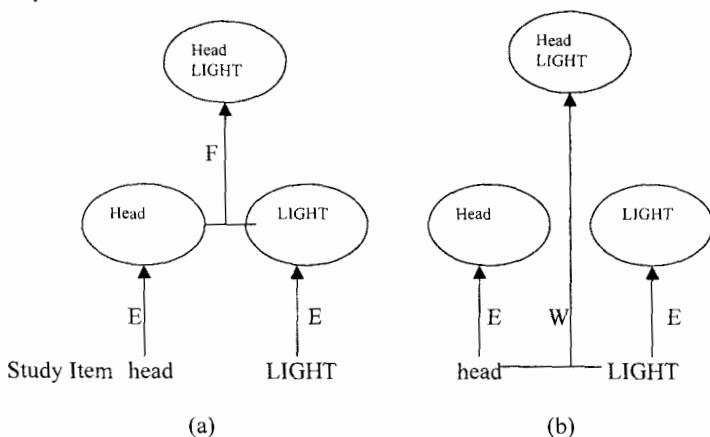


Figure 1. Two encoding hypotheses. In the single representation hypothesis the encoding of the representation of the entire study pair is dependent upon the encoding of its terms (a) but in the dual representation hypothesis the encoding of the representation of the entire study pair is independent of the encoding of its terms.

Hence, the target  $A_1B_1$  would not be discriminable from the distractors  $A_2B_3$  and  $A_3B_2$  on a subsequent recognition test. When  $E$  is less than 1.0 the critical assumption is that only part of the AB word pair may be encoded on a study trial—that is, either just the A word or just the B word. Flexser and Tulving (1978) call this the *goodness of encoding hypothesis*. It forms the basis of fragment theory, which has been used to make quantitative fits to data in a variety of recognition and recall tests (Jones, 1978, 1984; Ross & Bower, 1981).

Predictions of the goodness of encoding account were derived as follows. The strategy followed was to begin with the simplest possible model and to only add parameters if there were significant deviations between predictions and observations. So initially the value of  $F$ , the probability of an association forming between A and B given that they were both encoded, was set to 1.0 and hence dropped out of the equation for encoding the entire AB term, which became  $E^2$ . Let the probability that the AB representation of a study item was encoded be denoted by  $P_{ENC}(AB)$ . So  $P_{ENC}(AB) = E^2$ . Let the probability that the A and B terms of a study item were encoded separately be denoted by  $P_{ENC}(A,B)$ .

When  $F = 1$ ,  $P_{ENC}(A,B) = 0$ , because whenever a study item appeared its terms were encoded together in a single representation. So it is never the case that a participant remembers seeing the A term and remembers seeing the B term, but does not remember seeing the pair together. There may be a representation of A, or of B, or of the entire AB pair encoded in memory, but not distinct representations of both A and B. This restriction on the kinds of representations in memory only occurs for the single trace hypothesis illustrated in Figure 1(a) and only when  $F = 1$ . Recall that the encoding of only a single trace is called the trace identity assumption by Flexser and Tulving (1978) and is an essential assumption of their theory.

Similarly, the probability of encoding just A or B was  $2 * E * (1 - E)$ . Let the probability that just the A or B term was encoded be denoted by  $P_{ENC}(X)$ , where  $X = A$  or  $B$ . So  $P_{ENC}(X) = 2 * E * (1 - E)$ . The probability of encoding neither A nor B, denoted  $P_{ENC}()$ , was  $(1 - E)^2$ . All these values are collected in Appendix A.

Hence, during a recognition test we have four possible kinds of AB targets, categorized according to the level of representation that a target matches: whole (AB), both terms (A,B), one term (X), and neither term ( $\emptyset$ ). However, when  $F = 1$ ,  $P_{ENC}(A,B) = 0$ , so only three kinds of targets occur. We also have two kinds of B targets; those that have been encoded,  $P_{ENC}(B) = E$ , and those that have not been encoded,  $P_{ENC}(nB) = 1 - E$ . We need to consider the probability,  $P_{SE}$ , that each kind of target will be selected when it appears in the recognition task. We make the assumption that the more complete the match between the target and the representation in memory, the higher the probability that it will be selected as a target. So,  $P_{SE}(AB/AB) = S_4$ ,  $P_{SE}(AB/A,B) = S_3$ ,  $P_{SE}(AB/X) = S_2$ , and  $P_{SE}(AB/\emptyset) = S_1$ , where  $S_4 > S_3 > S_2 > S_1$ . Also, we let  $P_{SE}(B/B) = S_3$  and  $P_{SE}(B/\emptyset) = S_1$ . These probabilities are shown in Appendix A. Notice that  $S_1$  might plausibly be zero but it is included in order to provide a more general form of the model. The probability of recognizing each kind of target,  $P_{RN}$ , is the probability of it being encoded,  $P_{ENC}$ , times the probability of the target being selected,  $P_{SE}$ . These probabilities are shown in Appendix A. Notice that since  $F = 1.0$ ,  $P_{ENC}(A,B) = 0$ ,  $P_{ENC}(A,B) * P_{SE}(AB/A,B) = 0 * S_3 = 0$  regardless of the value of  $S_3$ . Nevertheless,  $P_{RN}(AB/A,B) = P_{ENC}(A,B) * P_{SE}(AB/A,B)$  is included for completeness and for ease of comparison with models for which  $P_{ENC}(A,B)$  is not equal to 0. Finally, the probability of recognizing a target is simply the sum of the probabilities of recognizing each kind of target, as shown in Appendix A.

The model presented here includes two assumptions of Flexser and Tulving's (1978) model: encoding variability and trace identity. However, the third key distinctive assumption of their model, retrieval independence, has not been incorporated. According to Flexser and Tulving the probability of a B target matching its representation will be partly independent of the AB target matching its representation because the target features compared in memory on different test trials are at least partly independent. That is, even if the entire AB study pair has been encoded in memory, the B target may activate some features in memory that are used in identifying it that are not subsequently activated by the AB target that contains it and vice versa. They call this disjunction between the features activated by the different targets retrieval independence. In contrast, the model here assumes that when the entire AB study pair has been encoded in memory the B target always activates a subset of the features that are subsequently activated by the AB target that contains it. So retrieval dependence rather than retrieval independence is assumed.

As mentioned above, the degree of independence between  $P(RnB)$  and  $P(RnAB)$  will be assessed with Yule's  $Q$ . As is described below, in order to compute Yule's  $Q$  it is necessary to first compute one of the cell frequencies—for example,  $P(RnB \cap RnAB)$ . To compute  $P(RnB \cap RnAB)$  we must specify the relationship between recognizing the AB target and encoding the B target. As shown in Appendix A,  $P(RnAB) = P_{RN}(AB/AB) + P_{RN}(AB/A,B) + P_{RN}(AB/X) + P_{RN}(AB/\emptyset)$ . According to the model, encoding the entire pair, AB, implies encoding B, encoding A and B implies encoding B, and encoding A or B implies encoding B on half of the trials. So  $P_{RN}(AB/AB) + P_{RN}(AB/A,B) + (1/2) * P_{RN}(AB/X)$  is the proportion of trials on which the AB target was recognized and the B target also matched its representation in memory. So  $(P_{RN}(AB/AB) + P_{RN}(AB/A,B) + (1/2) * P_{RN}(AB/X)) * P_{SE}(B/B)$  is the proportion of trials on which both the B target matches its representation in memory and is recognized and the AB target matches its representation in memory and is recognized. Conversely, if just A or neither A nor B is encoded then B is not encoded. So  $(1/2) * P_{RN}(AB/X) + P_{RN}(AB/\emptyset)$  is the proportion of trials on which the AB target was recognized and the B target did not match its

representation in memory. So  $((1/2)*P_{RN}(AB/X) + P_{RN}(AB/)) * P_{SE}(B/)$  is the proportion of trials on which both the B and AB targets are recognized though the B target does not match its representation in memory. Notice again that this value might plausibly be zero, but it is included for the purpose of generality. Hence,  $P(RnB \cap RnAB) = (P_{RN}(AB/AB) + P_{RN}(AB/A,B) + (1/2)*P_{RN}(AB/X)) * P_{SE}(B/B) + ((1/2)*P_{RN}(AB/X) + P_{RN}(AB/)) * P_{SE}(B/)$ . Notice also that the probability of recognizing a B target given that the corresponding AB target was also recognized,  $P(RnB | RnAB)$ , is equal to  $P(RnB \cap RnAB) / P(RnAB)$ , since this value will also be reported.

Both the predicted and the observed results for Gardiner's (1994) first two experiments are shown in Table 1. As can be seen from the table, the results of all four experimental conditions are predicted by the five-parameter model. Even though statistical tests of the fit of the model are not possible because Gardiner (1994) did not report the deviations of his means, only two of the predictions were not identical to the observed values, and the largest difference was .02. This fit was obtained despite the assumption that retrieval of the AB representation implied retrieval of the B representation. This is the opposite of the retrieval independence assumption. Instead the degree of independence observed between  $P(RnB)$  and  $P(RnAB)$  was modelled by assuming that finding a match between a probe and representation does not automatically mean the probe is selected as the target but only that it is selected with some probability.

## Confidence and the remember/know distinction

There is another reason why Gardiner (1994) may have failed to find a different level of independence in his successive recognition tests than in recognition followed by cued recall: Different response criteria in the tests may have masked the effects of cue overlap.

Gardiner (1994) did not have participants rate the confidence in their responses. Subsequently, Sikström and Gardiner (1997) made use of the dual-process theory of retrieval to explain similar results. According to this theory there are two dissociable retrieval processes that result in familiarity versus recollection judgement (Jacoby, 1991; Mandler, 1980). More recently, these have been characterized as know versus remember judgements (Gardiner & Java, 1993). Sikström and Gardiner found that the degree of dependence between recognition and subsequent cued recall of a B word was lower for know than for remember responses. Donaldson (1996) argued that the distinction between remembering and knowing is better understood as a distinction between responses that lie above a second decision criterion and responses that do not. Therefore, when responses in successive recognition tests are scored as correct only when they exceed a conservative decision criterion, test performance will be more dependent than if responses are scored according to a lax criterion. Whether or not remember and know responses reflect different retrieval processes, it is clear from Donaldson's analysis that employing a conservative criterion restricts the analysis to what participants would identify as remember responses. So extending the remember-know hypothesis to the successive recognition paradigm, the failure to find a higher level of dependence compared to the recognition-cued recall paradigm could have been the result of a higher number of low-confidence judgements in Gardiner's successive recognition experiments than in the recognition-cued recall paradigm. Suppose that the cued-recall test implicitly induced a conservative criterion for making a response, since a participant had to respond with a particular word, and participants implicitly adopted a "remember" criterion for this test. Further suppose that the AB

recognition test implicitly induced a lax criterion, since a participant implicitly adopted a "know" criterion to merely select a presented word-pair as a target. The laxer criterion in the AB recognition test might have counteracted the effect of increased cue-overlap and held the level of independence approximately constant.

## Purpose of experiments

The purpose of the experiments reported here was to test key assumptions of previous accounts of the degree of independence observed in successive recognition tests for an AB word-pair: retrieval independence and trace identity assumptions. In contrast, the new model presented above assumes the opposite of retrieval independence: retrieval consistency. It assumes that if a B target is presented on a recognition test, and an AB target is presented on a successive recognition test, and the AB target matches and retrieves an AB representation, then the probability that the B target previously matched and retrieved a B representation is 1.0. This assumption was rejected by both Flexser and Tulving (1978) and by Gardiner (1994). We shall see that despite our rejection of the retrieval independence/contextual hypothesis, the degree of independence observed between successive recognition tests can be accounted for by considering the effects of encoding variability and distractor context.

Three experiments were performed in the study reported here. Each experiment examined an assumption by varying the similarity between targets on successive recognition tests and the distractor contexts in which they appeared. In each experiment a recognition test for the B word was followed by a recognition test for the AB study pair. Confidence judgements were always collected. So it was possible to assess whether recognition independence was related to confidence and an artifact of a comparison between remember versus know judgements.

## EXPERIMENT 1

In applying the quantitative model to Gardiner's (1994) results a distinction was made between retrieval independence and postretrieval selection independence, since these terms referred to different parameters in the model. However, at this point it may seem to the reader that the distinction between retrieval independence and postretrieval selection independence is a distinction without a meaningful empirical or theoretical difference. It may also seem that when a model with five estimated parameters (or even four, if  $S_1$  is excluded) is used to predict three data points, even a perfect fit is not very impressive. To remedy these defects an approximate replication of Gardiner's experiment was performed and is reported below as Experiment 1. In the replication Gardiner's yes/no recognition test was replaced with a three-alternative forced-choice recognition test. This change made it possible to quantify the selection independence predictions with only a one-parameter model and to sharpen the distinction between retrieval independence and selection independence.

Our Experiment 1 differed from that of Gardiner's (1994) on a few points: All participants studied the whole set of 48 word pairs, and all participants were given a retention interval of 24 hours. Gardiner had found independence between recognition tasks at retention intervals of from 10 min to 4 days. One day was chosen in this study because a 10-min retention interval might be open to the objection that the results were a mixture of retrieval from long-term and short-term memory. The participants were randomly assigned to one of two groups that

differed with respect to time of presentation of study items. The 36 participants in one group studied the word pairs at a rate of 3 s per item, which Gardiner used in his first experiment, and the 36 participants in the other group studied the same word pairs at a rate of 5 s per item, which Gardiner used in his second experiment.

As mentioned above, forced-choice recognition tests were given. The first recognition test was for the B words alone in which each target was paired with two distractors that had not appeared on the study list. The second recognition test was for the AB pairs. Each target was paired with two distractor pairs made up of weakly associated words that did not appear on the study list.

## Method

### *Participants*

A total of 72 undergraduate students from the University of Oslo served as paid participants in the study. All participants were native speakers of Norwegian.

### *Materials and procedure*

The Norwegian translation (Lian, Glass, & Raanaas, 1998) of Tulving and Thomson's (1973) 48 word pairs (e.g., ground COLD) were used as study items. This is included as Appendix A.

Two equal groups of participants were given a study task in a lecture hall where word pairs were presented on overheads in a randomized order. One half of the participants studied the word pairs at a rate of 3 s per item, and the other half studied the same word pairs at a rate of 5 s per item.

The A word was printed in lower-case and the B word in upper-case letters, and the A word always appeared to the left of B. Participants were told to study each pair carefully in expectation of two memory tests the following day. The subsequent memory tests were mentioned to provide an incentive for the participants to take the study task seriously. However, no details beyond the phrase "two memory tests" were given, so the participants had no reason to adopt any particular study strategy.

A retention interval of approximately 24 hours elapsed between the study task and the tests. Participants were first given a B word recognition test. The test was in the form of a printed questionnaire that contained a cover page with instructions. In this test the 48 target words were mixed with 96 new words—that is, 144 items presented in 48 rows of 3. There were 16 rows per page. Each row contained one randomly positioned target word. The two distractors in each row were semantically related to the target—for example, SMOKETASTESMELL, where SMOKE was the target. Participants were told that each row contained one and only one target word from the study list. Furthermore, they were told to circle one word in each row and to respond to every row in the test in this way. They were also told to rate their degree of confidence by writing beside each row 3 for "certain", 2 for "may be", or 1 for "guessing". Each participant worked on the questionnaire independently and, for the most part, alone. That is, the day after the study session each participant returned to the laboratory and received a questionnaire that he or she filled out under the supervision of the experimenter. Sometimes more than one participant might have been working on a questionnaire at the same time but since they came and went individually there was no interaction among them.

All participants had to work no less than 10 and no more than 15 min on the task. After 2–3 min of "social talk" with the experimenter, participants were given the second test of recognition. This was another printed questionnaire. This test included 48 rows of word pairs, three word pairs in each row. Each AB target was paired with two distractor pairs made up of weakly associated words that did not appear on the study list. The participants were told that each row contained one of the word pairs that appeared in the study list the day before. Participants were then told to circle the one word pair they

thought had appeared as an AB item in the study list and rate their degree of confidence by writing 3 for "certain", 2 for "may be", or 1 for "guessing". Also, in this test a time limit of 15 min was imposed.

## Results

Two criteria were used for scoring the data. For the lax criterion correct responses were scored as hits regardless of the confidence rating given by the participant. For the conservative criterion only hits that were assigned a confidence rating of 3 were counted as hits.

Yule's  $Q$  was used to assess the degree of independence because it is a popular measure of correlation for  $2 \times 2$  contingency tables (Kahana, 2000). Let the results of the two recognition tests be  $P(Rn1)$ , the probability of recognition on Test 1, and  $P(Rn2)$ , the probability of recognition on Test 2. Let  $A = P(Rn1 \cap Rn2)$ , the proportion of items that are recognized on both tests,  $B = P(nRn1 \cap Rn2)$ , the proportion of items that are only recognized on the second test,  $C = P(Rn1 \cap nRn2)$ , the proportion of items that are only recognized on the first test, and  $D = P(nRn1 \cap nRn2)$ , the proportion of items that are recognized on neither test. Yule's  $Q$  is given by the equation:  $Q = (A*D - B*C)/(A*D + B*C)$ . Both a Yule's  $Q$  of  $-1.0$  and that of  $+1.0$  mean complete dependence, while a Yule's  $Q$  of  $0$  means complete independence between the tests compared.

The results are shown in Table 2. First consider the results when responses were scored according to the lax criterion. As can be seen from the table, there is even less dependence between the results of our recognition tests than in Gardiner's (1994) data. These unexpected results have not been anticipated by contemporary memory theory. According to Gardiner the informational overlap hypothesis predicts that there would be greater dependence between the successive recognition tasks (i.e., recognition of B and recognition of AB) than between recognition of the B word and subsequent recall of B given A as a cue, since only the former two tasks share a test cue.

The derivation of the parameters of the goodness of encoding/selection independence account was as shown in Appendix A, except for the selection of the target, which was now

TABLE 2  
Observed results and predicted results derived from three-parameter, dual representation, encoding variability model for Experiment 1

Time <sup>d</sup> per item	Criterion	Obs. Pre.	Yule's $Q$		$P(RnB)$		$P(RnAB)$		$P(RnB RnAB)$	$E$	$C$	$W$	
			$M$	$SD$	$M$	$SD$	$M$	$SD$					
3	Lax	Obs.	.26	.26	.57	.11	.78	.13	.60	.11			
		Pre.	.34		.54		.74		.60		.60	.34	.24
	Con.	Obs.	.51	.44	.12	.08	.43	.21	.23	.15			
		Pre.	.62		.23		.50		.32				
5	Lax	Obs.	.12	.54	.62	.09	.86	.10	.64	.10			
		Pre.	.17		.62		.86		.63		.89	.4	.36
	Con.	Obs.	.31	.55	.21	.10	.59	.23	.28	.13			
		Pre.	.18		.32		.65		.34				

Obs. = observed; Pre. = predicted; Con. = conservative.

<sup>a</sup>In s.

constrained by the forced-choice procedure so that  $P_{SE}(AB/AB) = P_{SE}(AB/A,B) = P_{SE}(B/B) = P_{SE}(AB/X) = 1$  and  $P_{SE}(AB/) = P_{SE}(B/) = 0$ . To assess the fit of the model, the observed and predicted values of three parameters were compared:  $P(RnB)$ ,  $P(RnAB)$ , and Yule's  $Q$ . Obviously the model should have fit all three values if it correctly described the degree of independence between  $P(RnB)$  and  $P(RnAB)$ . To evaluate the probability that the predicted and observed values were different,  $t$  scores were computed between the predicted and observed values as described by Hayes (1973, p. 399). A criterion of  $p = .05$  was set for these and all subsequent analyses.

For the lax criterion, a value of  $E = .36$  produced the best fitting predicted values for the 3-s condition of Yule's  $Q = .60$ ,  $t(35) = 7.74$ ,  $p < .05$ ;  $P(RnB) = .57$ ,  $t(35) = 0$ ,  $p > .05$ ; and  $P(RnAB) = .73$ ,  $t(35) = 2.27$ ,  $p < .05$ . A value of  $E = 0.43$  produced the best fitting predicted values for the 5 s condition of Yule's  $Q = 0.65$ ,  $t(35) = 5.81$ ,  $p < .05$ ,  $P(RnB) = 0.62$ ,  $t(35) = 0$ ,  $p > .05$ ; and  $P(RnAB) = 0.78$ ;  $t = 4.73$ ,  $p < .05$ . Since four out of six predicted data points were found to be different from the observed points at the  $p = 0.05$  level we may reject the one-parameter goodness of encoding model. The one-parameter model will not be considered further. In subsequent analyses not reported here varying the probability that both terms of a target were encoded as a single AB representation rather than distinct A and B representations (parameter F) did not change the best fitting parameter or fit of the model because, whether the target was encoded as a single or distinct representations, each was still selected over the distractors. So models in which parameter F was varied will also not be considered further.

In order to improve the fit of the model an implicit parameter was modified to move the model closer to reality. The one-parameter model assumed that an item encoded at study was always matched at test, and an item not encoded at study was never matched at test. The second part of the assumption, that an item not encoded at study would never be matched at test, cannot be correct. Consider a yes/no recognition test in which targets and distractors were presented one at a time, and a participant had to say whether each one was a target. If an item not encoded at study was never matched at test then the false alarm rate would always be zero. But this is not the case, as demonstrated by the seminal results for this task with these items reported by Tulving and Thomson (1973). Therefore a parameter C was added, which was the probability that a term that had not been presented at study would be matched at test. The changes in the derivation of the model when C is added are shown in Appendix B. All other parts of the derivation of the mathematical model are the same as those in Appendix A.

It was assumed that no forgetting occurred between the study task and the recognition test. The probability of a test item matching a representation during a recognition test,  $P_{MAT}$ , is shown in Appendix B. Each recognition trial contained two distractors. Each B distractor could either match or not match a representation for three different possible two-distractor contexts, designated  $P_{2M}(D)$ , as shown in Appendix B. For each AB distractor both, one, or neither term of the distractor might match a representation. In all there were six possible distractor contexts, as shown in Appendix B.

Hence, during a recognition test we have four kinds of AB targets, categorized according to the level of representation that a target matches: whole (AB), both terms (A,B), one term (X), and neither term ( ), and two kinds of B targets: (B) and (nB). We also have the different two-distractor-match contexts that a target can appear in. We need to predict the probability that each kind of target will be selected in each kind of two-distractor-match condition. In

order to predict these values the model must contain decision criteria that specify the kinds of target selected in each distractor context. The following four criteria were included in the model:

1. If an item matched an AB representation in memory then select that item.
2. If 1 is false, and both the A and B terms of one or more items match representations in memory, then select one of those items.
3. If 1 and 2 are false, and either the A or B term of one or more items matches a representation in memory, then select one of those items.
4. If 1, 2, and 3 are false then select any item.

These four criteria determine the probability of selecting each of the different kinds of target in the context of two distractors,  $P_{SE}(B|D2)$ , and so on, as shown in Appendix B. The probability of recognizing each kind of target,  $P_{RN}$ , is the probability of it occurring,  $P_{MAT}$ , times the probability of the target being selected,  $P_{SE}$ . These probabilities are also shown in Appendix B. Finally, the probability of recognizing a target is simply the sum of the probabilities of recognizing each kind of target, as shown in Appendix B.

The best fitting parameters and predicted data points of the two-parameter model were computed. For the lax criterion, for the 3-s condition the best fitting parameters of  $E = .73$  and  $C = .4$  generated predicted values of Yule's  $Q = .44$ ,  $t = 5.23$ ,  $p < .05$ ;  $P(RnB) = .57$ ,  $t(35) = 0$ ,  $p > .05$ , and  $P(RnAB) = .78$ ,  $t(35) = 0$ ,  $p > .05$ . For the 5-s condition the best fitting parameters of  $E = .82$  and  $C = .36$  generated predicted values of Yule's  $Q = .46$ ,  $t = 3.72$ ,  $p < .05$ ;  $P(RnB) = .62$ ,  $t(35) = 0$ ,  $p > .05$ , and  $P(RnAB) = 0.86$ ,  $t(35) = 0$ ,  $p > .05$ . For both Yule's  $Q$ s there was a significant difference between the predicted and observed values. So the two-parameter, encoding variability model can be rejected.

The single representation assumption shown in Figure 1(a) may be contrasted with a dual representation assumption shown in Figure 1(b). The single representation assumption is that the terms of the study pair are encoded either as a single representation or as distinct representations at study, but not as both a single representation and distinct representations. The dual representation assumption is that the terms of the study pair may be encoded both as a single representation and as distinction representations. For some items, such as head LIGHT and sun DAY, the dual representation assumption is especially plausible because the representation of the word pair may have an entirely different meaning from the representations of the individual words. It was assumed that the probability of encoding the single representation of the entire word pair was independent of the probability of encoding the distinct representations of the individual terms. So if a single representation of the whole pair (AB) was encoded with probability  $P_{ENC}(AB) = W$ , then the probability that distinct representations of the pair (A,B) were constructed but not a single representation of the entire pair was  $P_{ENC}(A,B) = (1 - W)*E^2$ . Similarly, the encoding probabilities for all the different fragments of the study terms were  $P_{ENC}(X) = (1 - W)*2*E*(1 - E)$  and  $P_{ENC}() = (1 - W)*(1 - E)^2$ . Also,  $P_{ENC}(B) = (1 - W)*E$  and  $P_{ENC}(nB) = (1 - W)*(1 - E)$ . Except for  $P(RnB \cap RnAB)$ , the rest of the derivation of the predicted observations was the same for the three-parameter model as for the two-parameter model shown in Appendix B. For the three-parameter model,  $P(RnB \cap RnAB) = P_{RN}(AB/AB)*P_{RN}(B/B) + (P_{RN}(AB/A,B) + (1/2)*P_{RN}(AB/X))*P_{SE}(B/B) + ((1/2)*P_{RN}(AB/X) + P_{RN}(AB/))*P_{SE}(B/)$ .

As can be seen from Table 2, the best fitting values of the three-parameter dual representation model are not significantly different from the observed values. For the lax criterion, the results of the  $t$  tests for the 3-s condition were for Yule's  $Q$ ,  $t(35) = 1.82$ ,  $p > .05$ ; for  $P(\text{RnB})$ ,  $t(35) = 1.61$ ,  $p > .05$ , and for  $P(\text{RnAB})$ ,  $t(35) = 1.82$ ,  $p > .05$ . The results of the  $t$ -tests for the 5-s condition were for Yule's  $Q$ ,  $t(35) = 1.42$ ,  $p > .05$ ; for  $P(\text{RnB})$ ,  $t(35) = 0$ ,  $p > .05$ , and for  $P(\text{RnAB})$ ,  $t(35) = 0$ ,  $p > .05$ . The predicted and observed values of  $P(\text{RnB} | \text{RnAB})$  are also shown because, as mentioned above,  $P(\text{RnB} | \text{RcB})$  has been ubiquitously reported in studies of the independence between recognition and cued recall, and a reader might want to compare the values. However, Reifer and Batchelder (1995) showed that this is not a good measure for model fitting, so no statistical tests are reported.

Now let us consider the results when items are scored according to the conservative criterion. When the responses are scored according to the conservative criterion, Yule's  $Q$  increases significantly for the 3-s,  $t(70) = 2.93$ ,  $p < .05$ , but not the 5-s,  $t(70) = 1.48$ ,  $p < .05$ , study condition. So there was more dependence between judgements for which the participants were certain. On the other hand, both values were significantly different from 1,  $t(35) = 6.59$ ,  $p < .05$ , and  $t(35) = 7.42$ ,  $p < .05$ , respectively, so there remained a significant amount of independence between  $P(\text{RnB})$  and  $P(\text{RnAB})$  to be explained.

Consider the fit of the three-parameter encoding variability model to the observations for conservative criterion judgements. Naturally, we must use the same parameter values that were used for the lax criterion because the conservative judgements are a subset of the lax ones. First consider the recognition of B targets. Obviously, in the case in which (1) the B target has been encoded and (2) neither distractor matches a representation a participant should be more confident than in other cases that do not contain this one. That is, the proportion of trials on which a confidence rating of 3 was assigned would be  $P'(\text{RnB}) = P_{\text{ENC}}(\text{B}) * P_{2\text{M}}(\text{D/nB;nB})$ . However, when this criterion was tried in the model it overpredicted the proportion of trials on which participants assigned a confidence rating of 3. A subset of these trials in which participants might be even more confident are those on which both the B target had been encoded as a distinct representation and the entire AB target had been encoded as a single representation. That is, the proportion of trials on which a confidence rating of 3 was assigned would be  $P'(\text{RnB}) = P_{\text{ENC}}(\text{B}) * P_{\text{ENC}}(\text{AB}) * P_{2\text{M}}(\text{D/nB;nB})$ . When this was tried it underpredicted the proportion of trials on which participants assigned a confidence rating of 3. Finally, the best fitting criterion was the larger subset of trials on which either both the B target had been encoded as a distinct representation and the entire AB target had been encoded as a single representation or both the A and B terms of the target had been encoded as distinct representations. So the probability of recognizing a B target with a confidence rating of 3 was  $P'(\text{RnB}) = (P_{\text{ENC}}(\text{B}) * P_{\text{ENC}}(\text{AB}) + P_{\text{ENC}}(\text{A,B})) * P_{2\text{M}}(\text{D/nB;nB})$ . However, as we shall see, this criterion again overpredicted  $P'(\text{RnB})$ .

Next consider the recognition of AB targets. Suppose that participants only assigned a confidence rating of 3 to judgements when (1) an AB target matched a memory representation of the entire AB pair or both terms of the AB pair and (2) neither distractor matched representations of both its terms. So the probability of recognizing the AB target with a confidence level of 3 was predicted to be  $P'(\text{RnAB}) = (P_{\text{ENC}}(\text{AB}) + P_{\text{ENC}}(\text{A,B})) * (P_{2\text{M}}(\text{D/X;X}) + (P_{2\text{M}}(\text{D/X;}) + P_{2\text{M}}(\text{D/;})))$ . Finally, consider the recognition of both a B target and its corresponding AB target. According to the criteria just outlined, just those AB recognition trials on which the AB target had been encoded as either a single representation or as two distinct

representations and the B target had appeared in the context of distractors that did not match their representations were the trials on which both B and AB were recognized according to the conservative criterion. So the probability of recognizing both the B target and the AB target with a confidence level of 3 was predicted to be  $P'(RnB \cap RnAB) = (P_{2M}(D/X;X) + P_{2M}(D/X;) + P_{2M}(D/;)) * P'(RnB) + P_{ENC}(A,B) * (P_{2M}(D/X;X) + P_{2M}(D/X;) + P_{2M}(D/;)) * P_{2M}(D/nB;nB)$ .

The predicted results are shown in Table 2. When the observed and predicted values were compared, for the 3-s condition, for Yule's  $Q$ ,  $t(35) = 1.48$ ,  $p > .05$ ; for  $P(RnB)$ ,  $t(35) = 8.13$ ,  $p < .05$ ; and for  $P(RnAB)$ ,  $t(35) = 1.97$ ,  $p > .05$ ; and for the 5-s condition, for Yule's  $Q$ ,  $t(35) = 0.32$ ,  $p > .05$ ; for  $P(RnB)$ ,  $t(35) = 4.73$ ,  $p < .05$ ; and for  $P(RnAB)$ ,  $t = 0.51$ ,  $p > .05$ . So the three-parameter model over-predicted the probability of a B target judgement receiving a confidence rating of 3 in the experiment.

## Discussion

So how are the results to be explained? Two post hoc explanations present themselves. First, the participant could select the correct AB pair in the second recognition test by recognizing the A term, the B term, or both, since none of these words appeared among the nontargets. Suppose that in the second task the decision about whether the AB pair was the target was largely or solely determined by whether the A word was remembered from the study list. Hence, in the first recognition test the judgement was based on the B word, because that was all that was available, and in the second recognition test, when the entire AB pair was present, the judgement was based on the A word. If we further assume that the similarity of the B word to its distractors in the first test was independent of the similarity of its associated A word to its distractors in the second test then we can explain the independence between the successive tests. This is essentially the explanation of the results proposed by Gardiner (1994), though he stated it in more general terms. In Gardiner's terminology the A term of the study pair is the *context*. This is why he called his explanation the contextual account.

A second explanation is the encoding variability/selection independence/dual representation model quantified above. It assumes that the entire AB word pair contains relational information defined by the presence of both words that is not available from either word alone. For example, the meaning of "head LIGHT" is not predictable from "head" and "LIGHT." If his relational information was the basis of judgements in the second recognition test and is encoded independently of the representations of the A and B terms then the independence between the results of the tests is explained by the model derived above.

These two explanations can be distinguished experimentally. Suppose we changed the second recognition test so that the distractors consisted of AB terms re-paired from the study list. That is, both the A and B terms of a distractor pair had appeared on the study list, but paired within a different item. In this case, a target pair on the AB recognition test could only be recognized if the participant recalled seeing the entire pair on the study list. If in fact participants are not encoding the specific pairs together, then the AB recognition test in which the distractors are re-paired AB items should be more difficult than the B recognition test in which the distractors were not shown on the study list.

Also, though the quantitative model derived from the encoding variability hypothesis provided an adequate description of the responses made under the lax criterion, it did not

provide an entirely adequate description of the responses made under the conservative criterion. To provide more data on this point confidence ratings were again collected.

## EXPERIMENT 2

The aim of the second experiment was to determine whether independence between successive part and whole recognition tests for the same target would be observed under conditions designed to require recognition of the entire AB target pair as an integrated unit.

This time, distractors in the second recognition test were created by re-pairing AB words that appeared in the study list. So on each test trial the A and B words of each pair had been seen on the study trial. However, only the target pair had been seen together. So if a participant recognized only the A or B word of a target this should have produced poorer performance on the AB recognition test than on the B recognition test in which the distractors had not been presented at study.

## Method

### *Subjects*

A total of 36 students (14 males and 22 females) from the University of Oslo served as paid participants. Mean age was 26 years with a range of 22–31 years. All participants were native speakers of Norwegian.

### *Materials and procedure*

The Norwegian version (Lian, Glass, & Raanaas, 1998) of the Tulving and Thomson (1973) 48 word pairs was used. The word pairs were presented, one at a time, on a PC screen. The presentation rate was set to one word every 5-s. The A word was printed in lower-case and the B word in upper-case letters, and the A word always appeared to the left of B. Participants were told to study each pair carefully in expectation of two memory tests the following day. The subsequent memory tests were mentioned to provide an incentive for the participants to take the study task seriously. However, no details beyond the phrase “two memory tests” were given, so the participants had no reason to adopt any particular study strategy. A retention interval of approximately 24 hours elapsed between the study task and the tests. Participants were first given a paper-and-pencil B word recognition test. The test questionnaire began with a cover page with instructions. In this test the 48 target words were mixed with 96 new words—that is, 144 items presented in 48 rows of 3. There were 16 rows on a page. Each row contained one randomly positioned target word. The two distractors in each row were semantically related to the target—for example, SMOKE TASTE SMELL, where SMOKE was the target. Participants were told that each row contained one and only one target word from the study list. Furthermore, they were told to circle one word in each row and respond to every row in the test in this way. They were also told to rate their degree of confidence by writing beside each row 3 for “certain”, 2 for “may be”, and 1 for “guessing”. All participants had to work no less than 10 and no more than 15 min on the task. Each participant worked on the questionnaire independently and, for the most part, alone. That is, the day after the study session each participant returned to the laboratory and received a questionnaire that he or she filled out under the supervision of the experimenter. Sometimes more than one participant might be working on a questionnaire at the same time but since they came and went individually there was no interaction among them.

After 2–3 min of “social talk” with the experimenter, participants were given the second paper-and-pencil test of recognition in the same format as that of the first. This test included 48 rows of word pairs,

three word pairs in each row. The participants were told that each row contained one of the word pairs that appeared in the study list the day before. The other two word pairs were formed by A and B terms that had appeared in the study list, but the combinations constituted new, nontarget word pairs—for example, ground—COLD head—GREEN cheese—LARGE, where ground—COLD was the target. Participants were then told to circle the one word pair they thought had appeared as an AB item in the study list and rate their degree of confidence by writing 3 for “certain”, 2 for “may be”, or 1 for “guessing”. Also, in this test a time limit of 15-min was imposed.

## Results

Under the lax criterion, every circled target was counted as a hit regardless of the confidence rating. Under the conservative criterion, correct responses were scored as correct only when given a confidence rating of 3. Table 3 summarizes the results according to the lax and conservative criteria.

Recall that in order to compute Yule's  $Q$  we must construct a  $2 \times 2$  contingency table whose cells are  $A = P(Rn1 \cap Rn2)$ ,  $B = P(nRn1 \cap Rn2)$ ,  $C = P(Rn1 \cap nRn2)$ , and  $D = P(nRn1 \cap nRn2)$ . Simpson's paradox is that if two or more  $2 \times 2$  contingency tables are collapsed into one, the summary table may show a relationship different from those shown by any of the original tables. Hintzman (1980) argued that Simpson's paradox must be considered in the analysis of memory retrieval. For example, suppose that in the present study some of the AB pairs were perfectly learned at study, the B term was always recognized on the first recognition test, and the entire AB term was always recognized on the second recognition test. So a  $2 \times 2$  contingency table for the two recognition tasks would show perfect learning and perfect dependence for these items. Further suppose that the remaining pairs were not learned at all. So performance was at chance in both recognition tests, and the  $2 \times 2$  contingency table for the two recognition tests would show no learning and perfect independence for these items. Nevertheless, when the two tables were combined in a single subjectwise analysis the contingencies would indicate both significant learning and a high degree of independence between the two recognition tests.

Due to Simpson's paradox, if individual items differ in recognizability then independence can be observed when the data are tabulated subjectwise (i.e., collapsed over items) even if there is complete dependence for each item. To rule out this explanation of the results separate

TABLE 3  
Probability of recognition in Experiment 2

Variable	Criterion						Parameters		
	Lax			Conservative					
	Obs.		Pre.	Obs.		Pre.	E	C	W
M	SD	Pre.	M	SD	Pre.				
P(RNB)	.59	.10	.59	.23	.13	.21	.65	.30	.67
P(RNAB)	.78	.14	.78	.45	.21	.49			
P(RNB RNAB)	.61	.10	.61	.30	.15	.36			
Yule's $Q$	.18	.50	.15	.34	.40	.46			

Obs. = observed; Pre. = predicted.

subjectwise and itemwise tabulations were performed on the data and analysed. In order to compare the results of these two tabulations the correlation between  $P(\text{RnB})$  and  $P(\text{RnB} | \text{RnAB})$  was computed as a second converging measure of independence for each tabulation. A correlation of 1.00 indicates complete independence. If high correlations in the subjectwise tabulations were the result of collapsing across items at different levels of difficulty then the correlations would be lower in the itemwise tabulations.

Consider first the results when the responses were scored according to a lax criterion. In the subjectwise tabulation of the data  $P(\text{RnB})$  was correlated  $r = .54$ ,  $CI\ 95\% = (.26, .74)$  with  $P(\text{RnAB})$ , and  $r = .92$ ,  $CI\ 95\% = (.85, .96)$  with  $P(\text{RnB} | \text{RnAB})$ . In a stepwise regression analysis with  $P(\text{RnB} | \text{RnAB})$  as the dependent and  $P(\text{RnB})$  and  $P(\text{RnAB})$  as the independent variables only  $P(\text{RnB})$  entered Equation 1, which explained 84% of the variance.

$$P(\text{RnB} | \text{RnAB}) = .05 + .92P(\text{RnB}) \quad 1$$

Confidence ratings in each of the recognition tests showed either a modest or a negligible correlation with  $P(\text{RnB} | \text{RnAB})$ . We also tabulated data itemwise and computed the correlation between  $P(\text{RnB})$  and  $P(\text{Rn} | \text{RnAB})$ , which turned out to be  $r = .95$ ,  $CI\ 95\% = (.90, .97)$ . Notice that, despite the correlation between  $P(\text{RnB})$  and  $P(\text{RnAB})$ , the correlation of .92 in the subjectwise analysis and .95 in the itemwise analysis between  $P(\text{RnB})$  and  $P(\text{RnB} | \text{RnAB})$  demonstrates almost complete independence between the two tests.

When the results were scored according to a conservative criterion, the correlation between  $P(\text{RnB})$  and  $P(\text{RnB} | \text{RnAB})$  was again computed from a subjectwise and an itemwise tabulation of the data. Subjectwise, the correlation was  $r = .85$ ,  $CI\ 95\% = (.73, .92)$ , and itemwise it was  $r = .85$ ,  $CI\ 95\% = (.74, .90)$ . As expected, the correlations between  $P(\text{RnB})$  and  $P(\text{RnB} | \text{RnAB})$  were lower when based on a conservative compared to a lax criterion. Yet, they remained at a remarkably high level in both the subjectwise and the itemwise analyses.

The similar levels of independence that were observed with both the subjectwise and itemwise tabulations greatly reduce the likelihood that the degree of independence observed was an artifact of Simpson's paradox. So quantitative predictions were derived from the encoding variability model and fit to the data. Because of the re-pairing of the A and B terms to create the distractors in the AB recognition test the values for the probability of a distractor matching a representation were  $P_{\text{MAT}}(D/A, B) = E^2$ ,  $P_{\text{MAT}}(D/X) = 2 * E * (1 - E)$ ,  $P_{\text{MAT}}(D/) = (1 - E)^2$ . The rest of the derivation of the predicted values is shown in Appendix B. The predictions are shown in Table 3. For the lax criterion, for Yule's  $Q$ ,  $t(35) = 0.33$ ,  $p < .05$ ; for  $P(\text{RnB})$ ,  $t(35) = 0$ ,  $p < .05$ ; and for  $P(\text{RnAB})$ ,  $t(35) = 0$ ,  $p < .05$ , so there were no significant differences between the predicted and observed values, and the model could not be rejected. For the conservative criterion it was again necessary to select the conditions to which a value of 3 was assigned. This time it was assumed that participants only assigned a confidence rating of 3 to B target judgements when (1) both the B target was encoded as a distinct representation, and the AB study item was encoded as a single representation and (2) on the test trial, neither distractor matched a representation. So the probability of recognizing a B target with a confidence rating of 3 was  $P'(\text{RnB}) = P_{\text{ENC}}(B) * P_{\text{ENC}}(AB) * P_{2M}(D/nB; nB)$ . Recall that this is a more restrictive criterion than the best fitting one for Experiment 1. We return to this point in the final discussion. For the recognition of AB targets the same criterion as in Experiment 1 was adopted even though the distractors were different. It was assumed that participants only assigned a confidence of 3 to judgements when (1) both a single representation of the entire AB

target had been encoded and in addition a distinct representation of at least one of its terms had been encoded and (2) not more than one distractor matched the separate representations of both its terms. That is, it was assumed that  $P'(RnAB) = P_{ENC}(AB) * (P_{ENC}(A,B) + P_{ENC}(X)) * (P_{2M}(D/X;X) + P_{2M}(D/X;) + P_{2M}(D/;))$ . Finally, consider the probability of recognition of a B target and its corresponding AB target. According to the criteria just outlined, if the AB target was recognized with a confidence level of 3 then, A, B or both A and B terms were also encoded. Furthermore, if the B term were encoded then its recognition would have been assigned a confidence level of 3 if it appeared in the context of two distractors that did not match their representations. So the predicted value of recognition of both the B target and the AB target was  $P'(RnB \cap RnAB) = P_{ENC}(AB) * P_{ENC}(B) * (P_{2M}(D/X;X) + P_{2M}(D/X;) + P_{2M}(D/;)) * P_{2M}(D/nB;nB)$ . The predicted values are shown in Table 3. When compared with observed values, for Yule's Q,  $t(35) = 1.63, p > .05$ ; for  $P(RnB)$ ,  $t(35) = 0.91, p > .05$ , for  $P(RnAB)$ ,  $t(35) = 1.97, p > .05$ . There was not a significant difference. So the three-parameter model could not be rejected.

## Discussion

To summarize, the results here confirmed and extended the results of Gardiner (1994) and of Experiment 1. When the responses are scored according to a lax criterion for both tests there was almost complete independence between recognition of the B item alone and recognition of the AB pair containing it. It is quite likely that over the set of all possible AB word pairs, pairs differ in the extent to which they are encoded as a single representation. Therefore, the degree of independence between tasks may be influenced by material specific properties when many different kinds of word pairs are examined. However, in this experiment correlations between  $P(RnB)$  and  $P(RnB | RnAB)$  were just as high or higher when based on an itemwise analysis than when based on a subjectwise analysis, which means that material specific properties played a minor role in the degree of dependence/independence observed between the tasks.

Also, Experiment 2 demonstrated for the first time that AB recognition is largely based on a single representation of the pair encoded at study. Otherwise the target could not be discriminated from distractors formed by re-pairing the terms of the study items.

The present results also show that the degree of dependence between the successive recognition tasks is influenced by the response criterion. A conservative criterion produced more dependence. However, the conservative criterion did not produce the degree of dependence that Sikström and Gardiner (1997) claim to be associated with conscious recollection of the encoding episode (remembering). Nor were the results under the conservative criterion adequately described by the encoding variability hypothesis.

## EXPERIMENT 3

The aim of the third experiment was to provide converging evidence that the AB recognition test was performed by accessing a single representation of the entire AB target. Experiment 2 consisted of the comparison of the triplet C B<sub>1</sub> D in the first recognition test and the comparison of the triplet A<sub>2</sub>B<sub>3</sub> A<sub>1</sub>B<sub>1</sub> A<sub>4</sub>B<sub>5</sub> on the second recognition test. Perhaps the inclusion of A<sub>2</sub>B<sub>3</sub> and A<sub>4</sub>B<sub>5</sub> distractors confused participants about which terms had been shown together at study and contributed to the level of independence that was observed. So Experiment 3

simplified the decision in the AB recognition test while leaving the derivation of the quantitative predictions of the encoding variability model unchanged.

Again the B target was compared with two words that had not appeared on the study list. This time, for half the participants the target word pair  $A_1-B_1$  was paired with the two distractors  $A_1-B_2$  and  $A_1-B_3$ . In this case, differences between A items were eliminated and could not form the basis for a response. For the remaining participants, when  $A_1-B_1$  was the target item  $A_2-B_1$  and  $A_3-B_1$  formed the distractors. In this case, differences between B items were eliminated and could not form the basis for a response. So each test was which opposite term appeared with  $B_1$  and  $A_1$ , respectively.

## Method

### *Subjects*

A total of 64 students from the University of Oslo served as paid participants in the experiment. Mean age was 24 years, range 20–40 years. All participants were native speakers of Norwegian. They were randomly assigned to one of two groups that differed with respect to the type of distractor items presented in the second recognition task. In Group I target and distractor items shared the same A word from the study list. In Group II target and distractor items shared the same B word from the study list.

### *Materials and procedure*

The Norwegian version of the Tulving and Thomson (1973) word pairs, used in Experiment 2, was also used in the present experiment. Groups of 10–12 participants were given the study task in a lecture room, where the word pairs were presented on overheads in a randomized order. The presentation rate was 5 s per word pair. The A and B word of a pair were printed in lower- and upper-case letters, respectively, and participants were told to study each pair carefully in expectation of two memory tests the following day.

A retention interval of approximately 24 hours elapsed between the study task and the tests. The first recognition test was exactly like the first test in Experiment 2. All participants had to work no less than 10 and no more than 15 min on the task. The second recognition test, which was presented immediately after the first test, differed for the two groups of participants. That is, the participants in Group I received a different printed questionnaire from the one given to the participants in Group 2. Both groups were given 48 rows of word pairs, three word pairs in each row. The 32 participants in Group I were told that each row contained one of the word pairs that appeared in the study list the day before. The other word pairs were recombinations of words from the study list. However, all three-word pairs included the same A word printed in lower-case letters—for example, head-COLD head-LIGHT head-LARGE. The participant was instructed to circle the one pair in each row that he or she thought was presented as a pair in the study task.

The 32 participants in Group II were also told that each row contained one pair that appeared in the study list the day before, and that the other two pairs were recombinations of words from the study list. However, all three pairs included the same B word printed in upper-case letters, and the participant was instructed to circle the one pair he or she thought had appeared as one pair in the study list—for example, ground-COLD head-COLD bath-COLD.

## Results

Table 4 summarizes the results of Experiment 3. The data show the same trend as the results of Experiment 2. When responses are scored according to a lax criterion there is almost complete retrieval independence between the two recognition tasks.

In Group I with the same A item in the second test and with responses tabulated subjectwise,  $P(\text{RnB})$  correlated  $r = .50$ ,  $CI 95\% = (.17, .73)$  with  $P(\text{RnAB})$  and  $r = .77$ ,  $CI 95\% = (.57, .88)$  with  $P(\text{RnBZ} | \text{RnAB})$ . In a stepwise regression analysis with  $P(\text{RnB} | \text{RnAB})$  as the dependent and  $P(\text{RnB})$  and  $P(\text{RnAB})$  as the independent variables only  $P(\text{RnB})$  enters Equation 2, which explains 59% of the variance.

$$P(\text{RnB} | \text{RnAB}) = .09 + .88P(\text{RnB}) \quad 2$$

Again, we made an itemwise analysis of the data. With a lax criterion we found that  $P(\text{RnB})$  correlated  $r = .92$ ,  $CI 95\% = (.87, .96)$  with  $P(\text{RnB} | \text{RnAB})$ , and with a conservative criterion we found that these variables correlated  $r = .88$ ,  $CI 95\% = (.79, .93)$ .

Similarly, in Group II when responses are scored according to a lax criterion,  $P(\text{RnB})$  correlated  $r = .59$ ,  $CI 95\% = (.31, .78)$  with  $P(\text{RnAB})$  and  $r = .91$ ,  $CI 95\% = (.83, .96)$  with  $P(\text{RnB} | \text{RnAB})$ . In a similar regression analysis for Group II, with the same variables, both  $P(\text{RnB})$  and  $P(\text{RnAB})$  enter Equation 3, which explains 88% of the variance.

$$P(\text{RnB} | \text{RnAB}) = .17 + .99P(\text{RnB}) - .18P(\text{RnAB}) \quad 3$$

In an itemwise analysis  $P(\text{RnB})$  correlated  $r = .96$ ,  $CI 95\% = (.93, .98)$  with  $P(\text{RnB} | \text{RnAB})$ . With a conservative criterion the itemwise correlation is  $r = .91$ ,  $CI 95\% = (.84, .95)$ . In both groups itemwise correlations between  $P(\text{Rn})$  and  $P(\text{RnB} | \text{RnAB})$  are higher than correlations between the same variables based on subjectwise tabulations. Also, correlations between  $P(\text{RnB})$  and  $P(\text{RnB} | \text{RnAB})$  with a lax criterion are stronger in Group II than in Group I. However, the correlation coefficients are not significantly different.

TABLE 4  
Probability of recognition in Experiment 3

Condition	Variable	Criterion						Parameters		
		Lax			Conservative					
		Obs.			Obs.			E	C	W
Same A item		M	SD	Pre.	M	SD	Pre.			
	P(RNB)	.56	.08	.56	.17	.08	.15	.62	.34	.57
	P(RNAB)	.71	.13	.74	.37	.19	.32			
	P(RNB   RNAB)	.58	.09	.58	.29	.13	.32			
Same B item	Yule's Q	.12	.45	.16	.47	.38	.54			
	P(RNB)	.57	.10	.55	.13	.08	.18	.60	.34	.69
	P(RNAB)	.79	.13	.79	.43	.21	.46			
	P(RNB   RNAB)	.60	.09	.59	.25	.14	.40			
	Yule's Q	.25	.29	.17	.51	.44	.47			

The Yule's  $Q$ s show significant increases in dependence from .12 to .47,  $t(62) = 2.59$ ,  $p < .05$ , and from .25 to .51,  $t(62) = 2.79$ ,  $p < .05$ , when the responses were scored according to a conservative criterion. Yet, these levels were still significantly below 1.0,  $t(31) = 7.77$ ,  $p < .05$ , and  $t(31) = 5.57$ ,  $p < .05$ , respectively.

The similar levels of independence that were observed with both the subjectwise and itemwise tabulations greatly reduce the likelihood that the degree of independence observed was an artifact of Simpson's paradox. So quantitative predictions were derived from the encoding variability model and fit to the data. The predictions are shown in Table 4. For the lax criterion, when the A term was held constant in the AB recognition test, for Yule's  $Q$ ,  $t(31) = 0.49$ ,  $p > .05$ , for P(RnB)  $t(31) = 0$ ,  $p > .05$ , and for P(RnAB),  $t(31) = 0$ ,  $p > .05$ ; and when the B-term was held constant, for Yule's  $Q$ ,  $t(31) = 1.15$ ,  $p > .05$ , for P(RnB),  $t(31) = 0$ ,  $p > .05$ , and for P(RnAB),  $t(31) = 0$ ,  $p > .05$ ; so there was not a significant difference between only a predicted and an observed value. For the conservative criterion, when the A-term was held constant in the AB recognition test, for Yule's  $Q$ ,  $t(31) = 1.03$ ,  $p > .05$ , for P(RnB)  $t(31) = 1.39$ ,  $p > .05$ , and for P(RnAB),  $t(31) = 1.47$ ,  $p > .05$ ; and when the B-term was held constant, for Yule's  $Q$ ,  $t(31) = 1.90$ ,  $p > .05$ , for P(RnB)  $t(31) = 3.48$ ,  $p < .05$ , and for P(RnAB),  $t(31) = 1.59$ ,  $p > .05$ . So, there was a significant difference between one pair of predicted and observed values for P(RnB).

## Discussion

Again partial independence was observed between successive recognition tests of the same study item. Since itemwise correlations between P(RnB) and P(RnB | RnAB) are higher than the corresponding subjectwise correlations, we can conclude that the degree of independence was not an artifact of different levels of recognition for different study items.

Let us consider again the two conditions in the experiment. In the condition in which the A word is held constant all three B words were targets, and in the condition in which the B word is held constant all three A words were targets. So if a participant did not have relational information about the specific AB pair seen during the study task performance in both these conditions would be at chance. That recognition of AB is significantly above chance demonstrates that the entire pair was encoded as a single representation. Furthermore, if the informational overlap account were correct then there should have been more independence between B recognition and AB recognition in which only the A term was varied than between B recognition and AB recognition in which only the B term was varied. In the former case there was no overlap between the cues, B versus A, necessary for discriminating the target from the distractors. But in the latter case there was complete overlap between the cues, B in both cases, necessary for discriminating the targets from the distractors.

## GENERAL DISCUSSION

The main empirical finding of the three experiments reported here was the significant level of independence observed between a B recognition test and a following AB recognition test for the same AB study item. Over three experiments, under a lax and a strict criterion, Yule's  $Q$  was always significantly different from 1.0, indicating a significant level of independence. In

fact, in one experiment Yule's  $Q$  was 0, so it was not significantly different from complete independence. Furthermore, in a second test of independence, the correlation between  $P(\text{RnB})$  and  $P(\text{RnB} | \text{RnAB})$  was as high in the itemwise analysis as in the subjectwise tabulation. So the significant level of independence cannot be attributed to an artifact of item differences in level of recognition. Finally, in Experiment 2 an  $A_1B_1$  target was discriminated from an  $A_2B_3$  and an  $A_4B_5$  distractor. In Experiment 3, an  $A_1B_1$  target was discriminated from an  $A_1B_2$  and an  $A_1B_3$  distractor. So in both experiments the target could only be discriminated from the distractors by matching it to a representation of the entire AB study item. So the independence between B and AB recognition cannot be because partly encoded AB targets were recognized on the basis of the A term alone.

There are several theoretical implications of this empirical finding. First the overall results under the lax criterion are well described by a three-parameter quantitative model derived from the encoding variability, selection independence, and dual representation assumptions. The encoding variability assumption is that on some study trials not all of the study item is encoded. The selection independence assumption is that when only part of the target has been encoded the probability of it being selected will depend on the distractor context it occurs in because a distractor may match a representation in memory as completely as the target. Hence, when distractors are selected independently in successive recognition tests there will be selection independence for incompletely encoded targets. The role of the distractors in determining a recognition judgement is not something explicitly considered by Flexser and Tulving (1978) or Gardiner (1994). The dual representation assumption is that a single representation of the entire AB study pair is created independently of distinct representations of each of its terms. This assumption explicitly contradicts Flexser and Tulving's trace identity assumption.

In short, the model presented here reverses two key assumptions of the informational overlap and contextual hypotheses in order to account for the degree of independence observed between successive recognition tests. In the former models, retrieval independence is the result of different targets sampling different features from the same representation because it is assumed that feature retrieval is probabilistic. In the model presented here retrieval independence is the result of different targets sampling different features from different representations because feature retrieval is consistent. No comparable quantitative model incorporating either the informational overlap or contextual assumption has been fitted to a similar set of recognition data.

Two findings are supportive of a single relational representation for the entire AB study item being encoded. In Experiments 2 and 3 the AB target could only be selected on the basis of a single representation of both terms. This result demonstrates a high degree of consistency in the features retrieved since distinguishing features of both terms had to be retrieved in order to select the target. Also, if independence were entirely the result of encoding variability then increasing study time should decrease encoding variability and hence increase dependence. In contrast, if independence is in part the result of encoding an independent relational representation for the entire AB study term then increasing study time should increase the probability of this representation being encoded and hence increase independence. A review of the three experiments indicates that Yule's  $Q$  was greater for the 3-s study condition than for any of the of the 5-s study conditions, which is consistent with an increased probability of encoding a relational representation of the entire study item.

If it is assumed that assigning a confidence level of 3 to a recognition judgement, reflecting certainty, is a remember judgement, while assigning a confidence level of 2 is a know judgement, then another issue addressed by these results is the remember-know distinction. The issue, as stated in Donaldson's (1996) critique of the remember-know hypothesis, is whether remember versus know judgements reflect qualitative or quantitative differences in the information on which the judgement was based. However, within the context of the model presented here this distinction disappears. Consider the case when a B term is recognized in the context of two distractors that do not match representations in memory. It may or may not be the case that the entire AB study term had previously been encoded. Suppose that this information is available during the recognition of the B term alone, and in the former case the judgement is called a remember response and in the latter case a know response. This distinction is consistent with a qualitative distinction between remember and know judgements. However, it is also consistent with Donaldson's hypothesis that remember judgements can be characterized as exceeding a higher criterion than that of know judgements. Within the framework of the dual representation hypothesis, high confidence remember judgements occur when a second representation is retrieved beyond the one actually matching the target. This qualitative difference is certainly consistent with Gardiner's (1994) hypothesis, but also results in a dual criterion quantitative model like the one proposed by Donaldson.

To summarize, the three-parameter model was used to predict 30 data points in five conditions over three experiments. Except for three data points, which were all P(RnB) under the conservative criterion, the predicted versus observed points were not significantly different. These results suggest that the assumptions of the model are correct.

However, a closer examination reveals that the three-parameter model cannot be the whole story. Since all three experiments used the same study task with the same items (with minor differences in study time) it is reasonable to assume that the probability of encoding a single representation of both terms of the representation would be consistent across all three experiments. Also, since in all experiments the B target recognition test was the same it is reasonable to assume that the same criterion for making confidence judgements would be used in all tests. However, a comparison of Table 2 with Tables 3 and 4 reveals that the probability of encoding the entire AB study pair was estimated to be much lower in Experiment 1 than in Experiments 2 and 3. Also, the criterion that predicted conservative confidence ratings for P(RnB) in Experiments 2 and 3 underpredicted the proportion of such ratings for Experiment 1. In fact, no criterion that was not completely post hoc predicted conservative confidence ratings for P(RnB) judgements in Experiment 1.

An important difference between Experiment 1 and Experiments 2 and 3 was the nature of the distractors presented in the AB recognition test. In Experiment 1 the distractors were made up of words that were not on the study list. In Experiments 2 and 3 the distractors were created by re-pairing the A and B terms of study items. Hence, in Experiments 2 and 3 the distractors were more similar to the targets than in Experiment 1, and required the retrieval of very specific information about which pair had appeared together on the study list in order to select the target. In Experiment 1, this very specific information about what had appeared together on the study list was not necessary to select the target. Suppose that in this case it either was not available or was not given determinant weight by the participant. Since this information was not being used as often in the AB recognition test it would lead to an underestimation of the proportion of trials that it was encoded on. This would produce the lower

estimates for parameter  $W$  in Experiment 1 than in Experiments 2 and 3. Furthermore, since the estimate of the proportion of conservative ratings for  $P(RnB)$  that predicted the results in Experiments 2 and 3 depended on the value of  $W$ , this value would also be underestimated in Experiment 1, which was the case.

The hypothesis that a more detailed representation of the target may only be available in the presence of more similar distractors is consistent with another well-established finding that until now has not received much comment. This is that forced choice recognition is higher when a target is paired with a more similar rather than a less similar distractor (Stewart & McAllister, 2001; Tulving, 1981). This result, together with the results here, suggests that recognition of a target may be based on a subset of the features retrieved from memory that are selected on the basis of the distractor context that it appears in.

If it is the case that a recognition judgement may be based on either representations of the individual terms of the study item or a single relational representation of the entire study item, depending on the distractor context, then it may well be the case that participants in Gardner's (1994) study did make use of distinctive features activated by the  $A$  term alone, as he suggested, to recognize the targets, though that was not sufficient for the experiments reported here.

Finally, research on  $B$  followed by  $AB$  recognition was originally motivated by research on  $B$  recognition followed by recall of  $B$  in response to  $A$ . In particular, it was specifically motivated by the hypothesis that there would be less independence in retrieval for the recognition tests than between recognition and recall. As we have seen, the opposite appears to be true. The higher level of independence between the recognition tests is consistent with another difference between the two paradigms. Årlemalm and Nilsson (1992) and Lian et al. (1998) found that the independence observed between the recognition and cued recall task was the result of the subset of items that were noun–adjective pairs. These words are shown in bold in Appendix C. They found that independence was observed in their studies because noun–adjective pairs tended to be recalled on  $A$ -cue recall trials regardless of whether the adjective was recognized on a  $B$  recognition trial. Hence, mean  $P(nRnB \cap RcAB)$ —that is, the probability that an item that is retrieved on the second test but not the first—was significantly greater for noun–adjective pairs than for other pairs. However, as indicated by the significant correlations between  $P(RnB)$  and  $P(RnAB)$  across all items, this was generally not the case in the successive recognition experiments reported here. This was confirmed by planned comparisons of noun–adjective pairs versus other items. However, in the 3-s presentation condition of Experiment 1 noun–adjective pairs had a mean  $P(nRnB \cap RnAB)$  of .44 whereas the remaining word pairs had a mean  $P(nRnB \cap RnAB)$  of .39,  $t(46) = 121$ ,  $p = .24$ . The 5-s presentation yielded mean  $P(nRnB \cap RnAB)$ 's of .44 and .32 for the noun–adjective pairs and the remaining word pairs, respectively,  $t(46) = 2.39$ ,  $p = .03$ . In Experiment 2  $P(nRnB \cap RnAB)$  was .42 for the noun–adjective pairs and .37 for the remaining pairs,  $t(46) = 1.17$ ,  $p = .25$ . In Experiment 3, for the same  $A$  distractors  $P(nRnB \cap RnAB)$  was .50 for the noun–adjective pairs and .38 for the remaining word pairs,  $t(46) = 2.11$ ,  $p = .04$ , and for the same  $B$  distractors  $P(nRnB \cap RnAB)$  was .41 for the noun–adjective pairs and .39 for the remaining word pairs,  $t(46) = 0.49$ ,  $p = .63$ . In all five groups of subjects the noun–adjective pairs always contributed a little more to independence than did the remaining word pairs. But this difference was significant only once or twice and certainly not large enough to explain the degree of independence observed over the entire item set.

So independence appears to be true of more items for successive recognition tests than for recognition followed by recall. Perhaps this is because independence between recognition and recall is restricted to those items for which there is a high probability of generating the B term from the A term by way of preexperiment association. Only future research can determine whether this hypothesis is true.

Finally, though this study was only concerned with recognition, it had implications for recall as well. Modern work on successive retrieval tests began with Tulving and Thomson's (1973) claim that recognition failure of recallable words was an interesting finding and that it undermined the generate and recognize explanation of recall. Their claim was based on the explicit assertion that  $P(\text{RnB}|\text{RnAB})$  would have been 1.00. Since this claim is false, some degree of recognition failure of recallable words necessarily follows from the recognition failure of recognizable words and is, of itself uninteresting. So it may be time to reconsider whether the generate and recognize remains the most satisfactory explanation of recall.

## REFERENCES

- Årlemalm T., & Nilsson, L.-G. (1992). Recognition failure of recallable words: Exception due to poor integration. *Scandinavian Journal of Psychology*, *33*, 266–276.
- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.
- Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*, 523–533.
- Flexser, 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., & Java, R. I. (1993). Recognition memory and awareness: An experiential approach. *European Journal of Cognitive Psychology*, *6*, 337–346.
- Hayes, W. L. (1973). *Statistics*. New York: Holt, Rinehart, & Winston.
- Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, *87*, 398–410.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.
- Jones, G. V. (1978). Recognition failure and dual mechanisms in recall. *Psychological Review*, *85*, 464–469.
- Jones, G. V. (1984). Fragment and schema models for recall. *Memory & Cognition*, *12*, 250–263.
- Kahana, M. J. (2000). Contingency analyses of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 59–72). New York: Oxford University Press.
- Lian, A., Glass, A. L., & Raanaas, R. K. (1998). Item specific effects in recognition failure: Reasons for rejection of the Tulving–Wiseman function. *Memory & Cognition*, *26*, 692–707.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252–271.
- 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.
- Reifer, D. M., & Batchelder, W. H. (1995). A multinomial modeling analysis of the recognition–failure paradigm. *Memory & Cognition*, *23*, 611–630.
- Ross, B. H., & Bower, G. H. (1981). Comparison of models of associative recall. *Memory & Cognition*, *9*, 1–16.
- Sikström, S. P., & Gardiner, J. M. (1997). Remembering, knowing and the Tulving–Wiseman Law. *European Journal of Cognitive Psychology*, *9*, 167–185.
- Stewart, H. A., & McAllister, H. A. (2001). One-at-a-time versus grouped presentation of mug book pictures: Some surprising results. *Journal of Applied Psychology*, *86*, 1300–1305.
- Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, *20*, 479–496.
- 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.

Original manuscript received 28 August 2001

Accepted revision received 23 July 2002

## APPENDIX A

Derivation of  $P(\text{RnB})$ ,  $P(\text{RnAB})$  and  $P(\text{RnB} \cap \text{RnAB})$  from encoding variability hypothesis for Gardiner's (1994) Experiments

*Probability of encoding entire study item (AB), both words of study item (A,B), one item of study item (X), and neither word of study item ( )*

$$P_{\text{ENC}}(\text{AB}) = E^2$$

$$P_{\text{ENC}}(\text{A,B}) = 0$$

$$P_{\text{ENC}}(\text{X}) = 2 * E * (1 - E), \text{ X} = \text{A or B}$$

$$P_{\text{ENC}}(\text{ }) = (1 - E)^2$$

Hence,

$$P_{\text{ENC}}(\text{B}) = P_{\text{ENC}}(\text{AB/AB}) + (1/2) * P_{\text{ENC}}(\text{AB/X}) = E$$

$$P_{\text{ENC}}(\text{nB}) = (1 - E)$$

*Probability of selecting targets that match complete and partial representations when paired with two distractors*

$$P_{\text{SE}}(\text{AB/AB}) = S_4$$

$$P_{\text{SE}}(\text{AB/A,B}) = P_{\text{SE}}(\text{B/B}) = S_3$$

$$P_{\text{SE}}(\text{AB/X}) = S_2$$

$$P_{\text{SE}}(\text{AB/ }) = P_{\text{SE}}(\text{B/ }) = S_1$$

*Probability of recognizing targets that match complete and partial representations*

$$P_{\text{RN}}(\text{B/B}) = P_{\text{ENC}}(\text{B}) * P_{\text{SE}}(\text{B/B})$$

$$P_{\text{RN}}(\text{B/nB}) = P_{\text{ENC}}(\text{nB}) * P_{\text{SE}}(\text{B/ })$$

$$P_{\text{RN}}(\text{AB/AB}) = P_{\text{ENC}}(\text{AB}) * P_{\text{SE}}(\text{AB/AB})$$

$$P_{\text{RN}}(\text{AB/A,B}) = P_{\text{ENC}}(\text{A,B}) * P_{\text{SE}}(\text{AB/A,B})$$

$$P_{\text{RN}}(\text{AB/X}) = P_{\text{ENC}}(\text{X}) * P_{\text{SE}}(\text{AB/X})$$

$$P_{\text{RN}}(\text{AB/ }) = P_{\text{ENC}}(\text{ }) * P_{\text{SE}}(\text{AB/ })$$

*Probability of recognizing a target*

$$P(\text{RnB}) = P_{\text{RN}}(\text{B/B}) + P_{\text{RN}}(\text{B/ })$$

$$P(\text{RnAB}) = P_{\text{RN}}(\text{AB/AB}) + P_{\text{RN}}(\text{AB/A,B}) + P_{\text{RN}}(\text{AB/X}) + P_{\text{RN}}(\text{AB/ })$$

$$P(\text{RnB} \cap \text{RnAB}) = (P_{\text{RN}}(\text{AB/AB}) + P_{\text{RN}}(\text{AB/A,B}) + (1/2) * P_{\text{RN}}(\text{AB/X})) * P_{\text{SE}}(\text{B/B}) + ((1/2) * P_{\text{RN}}(\text{AB/X}) + P_{\text{RN}}(\text{AB/})) * P_{\text{SE}}(\text{B/ })$$

## APPENDIX B

Derivation of  $P(\text{RnB})$ ,  $P(\text{RnAB})$  and  $P(\text{RnB} \cap \text{RnAB})$  for the two-parameter model of encoding variability hypothesis for Experiment 1

*Probability of a target (B;AB) and distractor (D) matching a representation during a recognition test*

$$\begin{aligned} P_{\text{MAT}}(\text{B}) &= P_{\text{ENC}}(\text{B}) \\ P_{\text{MAT}}(\text{AB}) &= P_{\text{ENC}}(\text{AB}) \\ P_{\text{MAT}}(\text{A,B}) &= P_{\text{ENC}}(\text{A,B}) + P_{\text{ENC}}(\text{X}) * C + P_{\text{ENC}}() * C^2 \\ P_{\text{MAT}}(\text{X}) &= P_{\text{ENC}}(\text{X}) * (1 - C) + P_{\text{ENC}}() * 2 * C * (1 - C) \\ P_{\text{MAT}}() &= P_{\text{ENC}}() * (1 - C)^2 \end{aligned}$$

$$\begin{aligned} P_{\text{MAT}}(\text{D/B}) &= C \\ P_{\text{MAT}}(\text{D/nB}) &= (1 - C) \\ P_{\text{MAT}}(\text{D/AB}) &= 0 \\ P_{\text{MAT}}(\text{D/A,B}) &= C^2 \\ P_{\text{MAT}}(\text{D/X}) &= 2 * C * (1 - C) \\ P_{\text{MAT}}(\text{D/}) &= (1 - C)^2 \end{aligned}$$

*Probability of two distractors matching representations on a recognition test trial*

$$\begin{aligned} P_{2\text{M}}(\text{D/B;B}) &= P_{\text{MAT}}(\text{D/B}) * P_{\text{MAT}}(\text{D/B}) \\ P_{2\text{M}}(\text{D/B;}) &= P_{\text{MAT}}(\text{D/B}) * P_{\text{MAT}}(\text{D/nB}) * 2 \\ P_{2\text{M}}(\text{D/nB;nB}) &= P_{\text{MAT}}(\text{D/nB}) * P_{\text{MAT}}(\text{D/nB}) \\ P_{2\text{M}}(\text{D/A,B;A,B}) &= P_{\text{MAT}}(\text{D/A,B}) * P_{\text{MAT}}(\text{D/A,B}) \\ P_{2\text{M}}(\text{D/A,B;X}) &= P_{\text{MAT}}(\text{D/A,B}) * P_{\text{MAT}}(\text{D/X}) * 2 \\ P_{2\text{M}}(\text{D/A,B;}) &= P_{\text{MAT}}(\text{D/A,B}) * P_{\text{MAT}}(\text{D/}) * 2 \\ P_{2\text{M}}(\text{D/X;X}) &= P_{\text{MAT}}(\text{D/X}) * P_{\text{MAT}}(\text{D/X}) \\ P_{2\text{M}}(\text{D/X;}) &= P_{\text{MAT}}(\text{D/X}) * P_{\text{MAT}}(\text{D/}) * 2 \\ P_{2\text{M}}(\text{D/;}) &= P_{\text{MAT}}(\text{D/}) * P_{\text{MAT}}(\text{D/}) \end{aligned}$$

*Probability of selecting targets that match complete and partial representations when paired with two distractors*

$$\begin{aligned} P_{\text{SE}}(\text{B} | \text{D2}) &= P_{2\text{M}}(\text{D/B;B}) * (1/3) + P_{2\text{M}}(\text{D/B;}) * (1/2) + P_{2\text{M}}(\text{D/nB;nB}) \\ P_{\text{SE}}(\text{nB} | \text{D2}) &= P_{2\text{M}}(\text{D/nB;nB}) * (1/3) \\ P_{\text{SE}}(\text{AB} | \text{D2}) &= 1 \\ P_{\text{SE}}(\text{A,B} | \text{D2}) &= P_{2\text{M}}(\text{D/A,B;A,B}) * (1/3) + P_{2\text{M}}(\text{D/A,B;X}) * (1/2) + \\ &P_{2\text{M}}(\text{D/A,B;}) * (1/2) + P_{2\text{M}}(\text{D/X;X}) + P_{2\text{M}}(\text{D/X;}) + P_{2\text{M}}(\text{D/;}) \\ P_{\text{SE}}(\text{X} | \text{D2}) &= P_{2\text{M}}(\text{D/X;X}) * (1/3) + P_{2\text{M}}(\text{D/X;}) * (1/2) + P_{2\text{M}}(\text{D/;}) \\ P_{\text{SE}}(/ | \text{D2}) &= P_{2\text{M}}(\text{D/;}) * (1/3) \end{aligned}$$

*Probability of recognizing targets that match complete and partial representations*

$$\begin{aligned} P_{\text{RN}}(\text{B/B}) &= P_{\text{MAT}}(\text{B/B}) * P_{\text{SE}}(\text{B} | \text{D2}) \\ P_{\text{RN}}(\text{B/}) &= P_{\text{MAT}}(\text{B/}) * P_{\text{SE}}(\text{nB} | \text{D2}) \\ P_{\text{RN}}(\text{AB/AB}) &= P_{\text{MAT}}(\text{AB/AB}) * P_{\text{SE}}(\text{AB} | \text{D2}) \\ P_{\text{RN}}(\text{AB/A,B}) &= P_{\text{MAT}}(\text{AB/A,B}) * P_{\text{SE}}(\text{A,B} | \text{D2}) \\ P_{\text{RN}}(\text{AB/X}) &= P_{\text{MAT}}(\text{AB/X}) * P_{\text{SE}}(\text{X} | \text{D2}) \\ P_{\text{RN}}(\text{AB/}) &= P_{\text{MAT}}(\text{AB/}) * P_{\text{SE}}(/ | \text{D2}) \end{aligned}$$

*Probability of recognizing a target for lax criterion*

$$\begin{aligned} P(\text{RnB}) &= P_{\text{RN}}(\text{B/B}) + P_{\text{RN}}(\text{B/}) \\ P(\text{RnAB}) &= P_{\text{RN}}(\text{AB/AB}) + P_{\text{RN}}(\text{AB/A,B}) + P_{\text{RN}}(\text{AB/X}) + P_{\text{RN}}(\text{AB/}) \\ P(\text{RnB} \cap \text{RnAB}) &= (P_{\text{RN}}(\text{AB/AB}) + P_{\text{RN}}(\text{AB/A,B}) + (1/2) * P_{\text{RN}}(\text{AB/X})) * P_{\text{SE}}(\text{B/B}) + ((1/2) * P_{\text{RN}}(\text{AB/X}) + \\ &P_{\text{RN}}(\text{AB/})) * P_{\text{SE}}(\text{B/}) \end{aligned}$$

*Probability of recognizing a target for conservative criterion*

$$\begin{aligned} P'(\text{RnB}) &= P_{\text{ENC}}(\text{AB}) * P_{2\text{M}}(\text{D/nB;nB}) \\ P'(\text{RnAB}) &= (P_{\text{MAT}}(\text{AB}) + P_{\text{MAT}}(\text{A,B})) * (P_{2\text{M}}(\text{D/X;X}) + (P_{2\text{M}}(\text{D/X;}) + P_{2\text{M}}(\text{D/;}))) \\ P'(\text{RnB} \cap \text{RnAB}) &= P_{\text{MAT}}(\text{AB}) * P_{2\text{M}}(\text{D/nB;nB}) \end{aligned}$$

## Appendix C

### English and Norwegian word pairs

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	2.56	NAIL TOUCH
11	memory-SLOW	hukommelse-TREG	2.14	QUICK START
12	covering-COAT	dekke-FRAKK	1.66	FUR CLOAK
13	barn-DIRTY	låve-SKITTE	1.42	SHINE POLISH
14	spider-BIRD	edderkopp-FUGL	1.38	SIGN FISH
15	crust-CAKE	skorpe-CAKE	2.00	BREAD BISQUIT
16	deep-SLEEP	dyp-SOVN	2.58	PILLOW HAMMOCK
17	train-BLACK	tog-SORT	1.32	CLOUD 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-PAIN	slå-SMERTE	2.94	FEELING EVIL
32	cloth-SUEEP	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	sol-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 TOOL
47	command-MAN	kommando-MANN	2.02	BOY GUY
48	fruit-FLOWER	frukt-BLOMST	2.34	BLOSSOM GRAIN