Memory & Cognition
1992, 20 (6), 601-611

A multinomial modeling analysis of the mnemonic benefits of bizarre imagery

DAVID M. RIEFER
California State University, San Bernardino, California

and

JEFFREY N. ROUDER
University of California, Irvine, California

A series of experiments was conducted to explore the cognitive processes that mediate the bizarreness effect, that is, the finding that bizarre or unusual imagery is recalled better than common imagery. In all experiments, subjects were presented with noun pairs that were embedded within bizarre or common sentences in a mixed-list design. None of the experiments produced a bizarreness effect for cued recall; however, for two of the experiments, the bizarre noun pairs were remembered significantly better than the common pairs for free recall. To determine if these differences were due to the storage or retrieval of the items, a multinomial model for the analysis of imagery mediation in paired-associate learning was developed and applied to the data from the experiments. The model revealed that bizarre sentences benefited the retrieval of the noun pairs but not their storage within memory. The empirical and modeling results are discussed relative to previous findings and theories on the bizarreness effect.

Many professional mnemonists (e.g., Lorayne & Lucas, 1974) advocate the use of unusual or bizarre imagery as a method for improving memory. The mnemonic benefit of bizarre over common imagery has been referred to as the bizarreness effect, and over the past two decades a number of experiments have been conducted to study it. However, the results of this research have been somewhat mixed. Most early research indicated that bizarre imagery is no more memorable than common imagery and that it may even have a negative effect (e.g., Bergfeld, Choate, & Kroll, 1982; Cox & Wollen, 1981; Wollen, Weber, & Lowry, 1972). But more recent research (e.g., Einstein, McDaniel, & Lackey, 1989; Hirshman, Whelley, & Palij, 1989) has shown that, under certain circumstances, bizarre imagery can have a reliable and beneficial effect on recall. For example, McDaniel and Einstein (1986) showed that the bizarreness effect occurs in mixed-list designs (with bizarre and common images both competing for recall) but not in between-list designs. Bizarre images are also more memorable than common images when they are tested for free recall (e.g., Pra Baldi, de Beni, Cornoldi, & Cavedon, 1985; Wollen & Cox, 1981) or delayed recall (Iaccino, Dvorak, & Coler, 1989; O'Brien & Wolford, 1982).

Now that the bizarreness effect is a reasonably well-established phenomenon, researchers have started to turn their attention to possible mechanisms or theories to account for this effect. Currently no single, unified theory exists that accounts for why bizarre imagery improves memory, or why it works under certain circumstances and not in others. Instead, Einstein and McDaniel (1987) have described more than a half dozen different theories that can be used to account for the bizarreness effect. A careful analysis of these theories reveals that many focus on storage versus retrieval explanations. For example, the attentional hypothesis (Merry, 1980; Wollen & Cox, 1981) states that bizarre images receive extra processing and are therefore stored better than common images. In contrast, the retrieval inhibition hypothesis (Hirshman et al., 1989; Einstein & McDaniel, 1987) postulates that bizarre images block the accessibility (or retrieval) of common images.

A key question, then, is whether the bizarreness effect is a storage or a retrieval phenomenon (or possibly is both). One approach to this issue is to conduct empirical experiments by using procedures designed to assess storage and retrieval separately. For example, a number of memory researchers have used the contrast between free recall, and either cued recall or recognition, to examine storage and retrieval (e.g., Hanley & Morris, 1987; Hogan & Kintsch, 1971; Schonfield & Robertson, 1966; Tulving & Psotka, 1971). Specifically, if a memory phenomenon occurs for free recall but not for cued recall or recognition, this is often taken as evidence for a retrieval explanation of the phenomenon because cued recall and recognition presumably lessen or eliminate the importance of retrieval processes. This pattern of results does in fact occur over many experiments on bizarre imagery. Ein-

This research was supported by NSF Grant BNS-8910552 to William Batchelder and D.M.R. We thank William Batchelder, Gilles Einstein, and two anonymous reviewers for their helpful comments on this manuscript. Requests for reprints should be sent to David M. Riefer, Department of Psychology, California State University, San Bernardino, CA 92407.

stein and McDaniel (1987), in an excellent review, have pointed out that, at least for immediate recall, the bizarreness effect is strongest for free recall but is not obtained for either cued recall (e.g., Andreoff & Yarmey, 1976; Webber & Marshall, 1978) or recognition (Emmerich & Ackerman, 1979; McDaniel & Einstein, 1986).

In addition, McDaniel and Einstein (1991) and Einstein et al. (1989), in a series of experiments on sentence recall, examined the effects of bizarre imagery on two separate measures: the number of sentences accessed during recall and the number of items recalled from accessed sentences. They showed that bizarre imagery improves the first measure but not the second. These two empirical statistics are analogous to measures developed by Tulving and Pearlstone (1966) to examine the availability (storage) and accessibility (retrieval) of words that come from taxonomic categories. Specifically, Tulving and Pearlstone used *P*(cat), the probability of recalling at least one item from a category, as a measure of retrieval processes and *IPC*, the mean number of items recalled per category, as a measure of storage. If we assume that a sentence and the elements within it comprise a "category," then number of sentences accessed corresponds to *P*(cat) and number of items recalled from accessed sentences corresponds to *IPC*. Thus, by extending Tulving and Pearlstone's logic to this situation, one can conclude, as did Einstein et al. (1989), that bizarre imagery improves the accessibility but not the availability of items.

Overall, the above pattern of results suggests that bizarre imagery aids the retrieval of items from memory but does not necessarily benefit their storage. However, a number of researchers have criticized the use of empirical statistics such as the ones above for separating storage and retrieval factors. For example, some theorists (Craik & McDowd, 1987; McNulty & Caird, 1966) have pointed out the methodological problems in comparing recall and recognition performance. A detailed criticism has been given by Smith (1980), who argues that examining the differences in free recall versus cued recall or recognition does not necessarily allow one to rule out entirely encoding processes as a factor in recall. This is because successful retrieval inherently depends on a sufficient level of encoding, which might not always be equivalent across groups or experimental conditions. Another line of criticism comes from Riefer and Batchelder (1988, 1991a), who argue that empirical statistics such as *P*(cat) and *IPC* can present a misleading picture of cognitive processing because they often result from a combination of different cognitive activities. In other words, a statistic designed to measure one specific cognitive process can sometimes be influenced by many different cognitive functions and thus may not be a pure measure of any single one. As an alternative to empirical statistics, Riefer and Batchelder (1988) have advocated a type of mathematical modeling, called multinomial modeling, for studying underlying cognitive processes. Multinomial models are relatively simple models that can be used as data-analysis tools for measuring unobservable cognitive events, and as such provide a more theoretically motivated measure of cognitive processing than do ad hoc statistics.

One example of this type of modeling, relevant to the current issue, is a model developed by Batchelder and Riefer (1980, 1986) for measuring storage and retrieval. Unfortunately, this model is inappropriate for studying the bizarreness effect, because bizarre-imagery experiments typically employ a paired-associate paradigm involving word pairs or sentences, whereas the storage-retrieval model analyzes category clustering data from a serial list-learning task. However, it should be possible to develop a new multinomial model specifically for this situation. The basic requirement for the development of a multinomial model is an experimental paradigm that generates data in several discrete categories. Paired-associate learning certainly provides such a paradigm, especially if the cued recall of paired associates can be supplemented with some other measure, such as free recall.

Thus, the purpose of this article is to present a new multinomial model for the analysis of imagery mediation in paired-association learning and to apply the model to the issue of bizarre imagery. The model examines the free and cued recall of word pairs, and uses these data to measure separately the storage and retrieval processes involved in their recall. The next section describes in detail the experimental paradigm behind the model and develops the model itself. Following this, the model is applied to a series of experiments on bizarre imagery, in an attempt to test the validity of the model and explore the storage-retrieval basis of the bizarreness effect.

MULTINOMIAL MODEL

Data Representation

The experimental paradigm behind the model involves a paired-associate learning task, which traditionally is one of the methods used by researchers to study bizarre imagery (e.g., Cornoldi, Cavedon, De Beni, & Pra Baldi, 1988; Iaccino & Sowa, 1989). In such a task, subjects memorize word pairs consisting of a stimulus term (S) and response term (R). During test the subjects receive the stimulus terms as cues to recall and are required to provide the response term for each stimulus. The trick is to supplement these cued-recall data with a free-recall task, in which subjects recall the word pairs without recall aids. Having subjects engage in both free and cued recall is a method used in previous studies on bizarre imagery (Hirshman et al., 1989; Pra Baldi et al., 1985; Wollen & Cox, 1981), as well as in other studies (e.g., Hirshman, 1988; Tulving & Pearlstone, 1966).

On free recall, subjects can recall both words (SR), just the stimulus term or response term but not both (S/R), or neither item (Ø). On cued recall, with the stimulus term provided, subjects can either recall the response (R) or not recall it (Ø). The combination of these possibilities creates six separate recall events: E_1—both items freely recalled, correct cued recall; E_2—one and only one item freely recalled, correct cued recall; E_3—neither item freely

recalled, correct cued recall; E_4—both items freely recalled, incorrect cued recall; E_5—one and only one item freely recalled, incorrect cued recall; and E_6—neither item freely recalled, incorrect cued recall. Furthermore, let N_i be the frequency of occurrence for event E_i, with $N = \sum_{i=1}^{6} N_i$.

Model Development

The multinomial model assumes that subjects' free- and cued-recall responses are a function of five hypothetical, dichotomous events:

Storing the stimulus-response association. During study, storage of the stimulus-response pair occurs when an adequate representation of the stimulus is encoded within memory and the response term is associated to the stimulus. Define a as the probability of forming and storing a stimulus-response association, $0 \leq a \leq 1$.

Retrieval of the association during free recall. During the free-recall task, the model assumes that a stored association is retrieved with probability r_1 $(0 \leq r_1 \leq 1)$. If this happens, both the stimulus and response terms are recalled.

Retrieval of the association during cued recall. If a paired associate is stored within memory, then the model assumes that the response term is retrieved during cued recall with probability r_2 $(0 \leq r_2 \leq 1)$.

Recall of unretrieved associates. If an associated word pair is not retrieved as a whole unit during free recall, then it is possible that exactly one of the items in a pair is recalled independently during free recall. This occurs with probability s_1 $(0 \leq s_1 \leq 1)$, and for simplicity, the model assumes that it is equally likely for the singleton to be the stimulus or response term.

Recall of nonassociated items. If a word pair is not stored as an associate within memory, then one of the items can be stored and retrieved as a singleton. This is analagous to the singleton recall of nonretrieved associates and occurs with probability s_2 $(0 < s_2 \leq 1)$. Again, the model assumes that recall of a singleton is equally likely to be a stimulus or response term.

Of these five parameters, the two that are of most interest are parameter a, which measures the storage of the paired associates, and parameter r_1, which measures their retrieval. Parameters r_2, s_1, and s_2 are less central to the storage-retrieval issue and can be regarded as "nuisance parameters" (Riefer & Batchelder, 1988). Parameters s_1 and s_2 represent the probability of storing and retrieving items as singletons from nonassociated or nonretrieved pairs, respectively, and thus combine both storage and retrieval factors. Parameter r_2, which in part is the probability of correct cued recall given correct free recall, is a necessary parameter in the model because occasionally subjects are correct on free recall but incorrect on cued recall for a given paired associate. This should be a fairly rare event, however, so a mild validity check on the model is that the estimate of r_2 should be close to 1 for all conditions.

In multinomial models such as this one, in which the recall events are assumed to be a function of all-or-none cognitive processes, it is often convenient to express the model in the form of a tree diagram (Riefer & Batchelder, 1988). Figure 1 presents the tree diagram for the current model, and from this tree structure it is easy to write expressions for the probabilities of each data event:

$$P(E_1) = ar_1r_2, \tag{1a}$$

$$P(E_2) = a(1-r_1)r_2s_1, \tag{1b}$$

$$P(E_3) = a(1-r_1)r_2(1-s_1), \tag{1c}$$

$$P(E_4) = ar_1(1-r_2), \tag{1d}$$

$$P(E_5) = a(1-r_1)(1-r_2)s_1 + (1-a)s_2, \tag{1e}$$

$$P(E_6) = a(1-r_1)(1-r_2)(1-s_1) + (1-a)(1-s_2). \tag{1f}$$

Parameter Estimation

The expressions in Equation 1 can be used to derive the likelihood function for the model, which is an equation that expresses the probability of the data as a function of the parameter values. For the multinomial model described above, the function is

$$\begin{aligned} L = {} & \left(\frac{N!}{N_1!N_2!N_3!N_4!N_5!N_6!}\right)[ar_1r_2]^{N_1}[a(1-r_1)r_2s_1]^{N_2} \\ & \times[a(1-r_1)r_2(1-s_1)]^{N_3}\,[ar_1(1-r_2)]^{N_4} \\ & \times[a(1-r_1)(1-r_2)s_1 + (1-a)s_2]^{N_5} \\ & \times[a(1-r_1)(1-r_2)(1-s_1) + (1-a)(1-s_2)]^{N_6}. \end{aligned} \tag{2}$$

The importance of the likelihood function is that it can be used to obtain maximum likelihood estimators (MLEs) for the parameter values. By definition, these are the values of the parameters that maximize the likelihood function. Fortunately, by using calculus methods (see

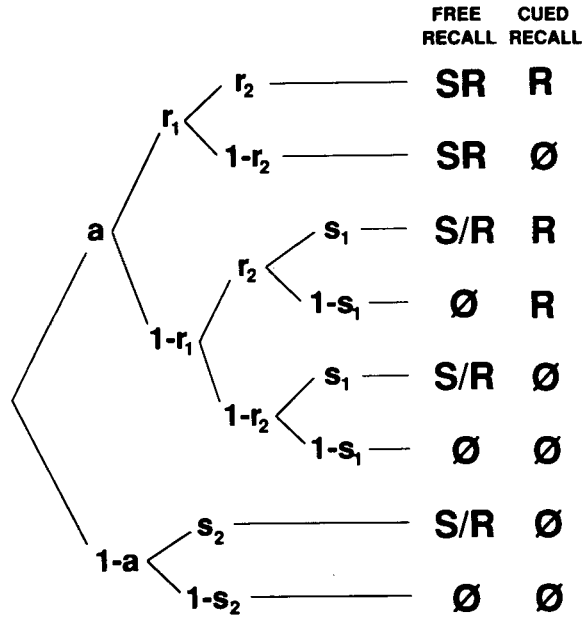

Figure 1. Tree structure for the multinomial model.

Hogg & Craig, 1978, chap. 6), it is possible to derive closed-form solutions for these parameter estimators:

$$a = [(N_1+N_2+N_3)(N_1+N_4)]/NN_1, \quad (3a)$$

$$r_1 = N_1/(N_1+N_2+N_3), \quad (3b)$$

$$r_2 = N_1/(N_1+N_4), \quad (3c)$$

$$s_1 = N_2/(N_2+N_3), \quad (3d)$$

$$s_2 = (N_1N_5-N_2N_4)/[(N_1(N_5+N_6)-N_4(N_2+N_3)]. \quad (3e)$$

The expressions in Equation 3 yield MLEs within the [0,1] interval, provided that the inequalities $N_1N_6 > N_3N_4$ and $N_1N_5 > N_2N_4$ both hold for the data set. If these inequalities do not hold, Riefer and Batchelder (1988) describe how iterative search procedures can be used to obtain MLEs for the parameter values within the unit interval.

One important feature of this model is that the number of parameters exactly equals the number of independent data events—five in each case. Consequently there are no degrees of freedom (*df*) left over for testing the goodness of fit for the model. This is not a problem for measurement because the model is identifiable and one can still derive unique parameter estimates. But it would be desirable to develop some restriction in the model's parameters to free up at least 1 *df* for testing the model's fit to data. One possible restriction is to set the two singleton parameters equal to each other, that is, $s_1 = s_2 = s$. This assumption, of course, will not be valid for all data sets because s_1 and s_2 arise from somewhat different cognitive events. However, it may be reasonable under some circumstances to constrain these parameters to be equal since they both reflect the process of individually recalling items that were not recalled as associated word pairs. This would present an opportunity to test the fit of the model, as well as the adequacy of this assumption itself.

We will refer to the unrestricted version of the model as Case I, and the restricted version with $s_1 = s_2 = s$ as Case II. The likelihood function for Case II of the model is similar to the one shown in Equation 2, except with s substituted for s_1 and s_2. Again, the MLEs for this version of the model are those values that maximize this new likelihood function, and fortunately it is again possible to derive closed-form solutions for these estimates. In fact, the estimates for a, r_1, and r_2 are unchanged and are exactly as they appear in Equations 3a-3c. The new MLE for the combined parameter s becomes

$$s = (N_2+N_5)/(N_2+N_3+N_5+N_6). \quad (4)$$

These new equations yield MLEs within the unit interval, provided the inequality $N_1(N_5+N_6) > N_4(N_2+N_3)$ holds for the data set.

It is important to note that the MLEs in Equations 3 and 4 are only approximate because they are based on the assumption that each subject has the same underlying parameter values. Although it is possible to obtain MLEs separately for individual subjects and average them, such estimates are highly variable and may be biased, especially if some of the N_i frequencies equal zero. It is thus common practice to sum the N_i statistics across subjects to avoid this problem. However, the use of group aggregate data to estimate the model's parameters ignores the possible effects of individual differences. This raises the issue of what effect small sample sizes and individual differences have on the accuracy of the MLEs, because such factors may lead to systematic bias in point estimation and to inflated confidence intervals.

As a way of addressing this issue, Riefer and Batchelder (1991b) advocate the use of Monte Carlo computer simulations to test the pre-asymptotic statistical properties of multinomial models. This is especially feasible because multinomial models are mathematically simple and easily simulated on personal computers. To explore this with the current model, we conducted a series of such simulations for a range of parameter values and sample sizes. Similar to Riefer and Batchelder's (1991b) approach, we used the beta distribution to simulate individual differences in the parameter values, with a standard deviation of 0 (no individual differences) or .1 (moderate individual differences—see Riefer & Batchelder, 1991b). We also aggregated the N_i statistics across subjects for each simulation. In general, the results of the simulations were encouraging, revealing that the model is quite robust in the presence of individual differences. Even for small sample sizes ($N = 50$) and moderate individual differences, the computed storage and retrieval parameters appeared to be unbiased and deviated from their true values by no more than 2%. Moreover, the variability of the parameter estimates was reasonably small when there were no individual differences. The standard deviation for any parameter never exceeded .07 for samples as small as 50 and, as expected, standard deviations decreased by the square root of the sample size. The introduction of individual differences roughly doubled the size of these standard deviations. These results imply that, for reasonably sized samples, researchers need not worry excessively about moderate individual differences in aggregating the N_i statistics across subjects for this model.

In summary, by collecting the N_i statistics for a given experiment, the multinomial model can theoretically measure storage and retrieval processes by computing estimates for the parameters a and r_1. Moreover, if different values of these parameters have been computed across a number of experimental conditions, then Riefer and Batchelder (1988) show how to conduct hypothesis tests, using the log likelihood ratio statistic G^2, to determine if the differences between parameters are statistically significant. We illustrate these techniques for parameter estimation and hypothesis testing in the next section, in which we apply the model to three experiments on bizarre imagery.

EXPERIMENTS 1, 2, AND 3

To test the multinomial model and explore storage and retrieval factors mediating the bizarreness effect, we con-

ducted three experiments on bizarre imagery. All three were basically replications of the same experimental design, except that slightly different stimuli were used in Experiment 3 than were used in Experiments 1 and 2. The reason for conducting three replications of the same experiment was to determine the consistency and reliability of the model's analysis.

The experiments were similar to recent studies on the bizarreness effect (e.g., Einstein et al., 1989; Hirshman et al., 1989; Kroll & Tu, 1988). Specifically, subjects were presented both bizarre and common stimuli in a mixed-list design using an incidental learning procedure. The stimuli consisted of sentences that described either a bizarre or common relationship between two objects. The subjects attempted to form a mental image as they read each sentence and rated the vividness of their images. This was followed by a surpise memory test involving first a free recall and then a cued recall of the stimulus pairs. On the basis of the results of prior research (Hirshman et al., 1989; Pra Baldi et al., 1985), we would expect to find a reliable bizarreness effect for free recall but not necessarily for cued recall. Moreover, the multinomial model should reveal whether any empirical effects of bizarre imagery are due to storage or retrieval.

Method

Subjects. The subjects were 95 undergraduate students from the University of California at Irvine, with 33 subjects in Experiment 1, 32 in Experiment 2, and 30 in Experiment 3. All were given extra course credit for their participation.

Materials. The stimuli for each experiment consisted of 12 word pairs, each composed of two concrete nouns (e.g., MINISTER-BIBLE). A common and bizarre sentence was constructed for each of these 12 noun pairs. For example, the common sentence for the noun pair MINISTER-BIBLE was "The MINISTER read the BIBLE," whereas the corresponding bizarre sentence was "The MINISTER ate the BIBLE." As can be seen in these examples, the 2 concrete nouns were capitalized within each sentence. A number of the sentences were modified from ones used by McDaniel and Einstein (1986). The same set of 12 noun pairs was used for Experiments 1 and 2. For Experiment 3, 4 of these pairs were replaced with new ones, because the subjects had difficulty forming a distinctive image or confused some of the nouns from different sentences (e.g., LAWYER and BANKER). A complete list of the sentences used in each experiment is given in the Appendix.

We pretested the stimuli used in Experiments 1 and 2 on 11 people, who rated each sentence on separate 5-point scales for bizarreness and level of interaction between the two objects (cf. Bergfeld et al., 1982; Wollen et al., 1972). As expected, the 12 bizarre sentences were rated significantly higher for bizarreness than were the 12 common sentences [M = 4.35 vs. 1.68, respectively; $t(11)$ = 10.24, $p < .001$]. However, bizarre sentences did not significantly differ from common sentences in their degree of interactiveness [M = 3.51 vs. 3.32, respectively; $t(11) = 0.72$, n.s.]. We obtained similar results when the 12 sentences from Experiment 3 were separately tested on 17 additional subjects.

Procedure. The subjects participated in groups ranging from 5-10 persons. Each subject sat at an IBM PC that presented the instructions and the stimuli. The subjects were instructed to form a mental image for each sentence and to rate the vividness of that image on a 5-point scale, with 1 = not vivid and 5 = very vivid. We chose the vividness rating task because McDaniel and Einstein (1986) have shown that the strongest bizarreness effect is obtained with this type of rating task rather than with other rating tasks. The subjects received 1 practice sentence to rate before they were presented with the 12 experimental sentences. For each subject, 6 of the experimental sentences were randomly chosen to be common, and 6 were chosen to be bizarre. Presentation order of the 12 sentences was also determined randomly for each subject. Each sentence appeared on the screen for 7 sec, during which time the subjects attempted to form their mental image. This was followed by a 3-sec period, during which they entered their vividness rating into the computer.

Following presentation of the 12 sentences, the subjects received a distractor task for 2 min, during which they circled small differences between pairs of almost-identical pictures. The purpose of the distractor task was to help eliminate recency effects in the recall of the stimuli. After this, the subjects received two surprise memory tests. The first was a free-recall test in which they recalled the noun pairs in writing on a blank sheet of paper. The subjects were instructed to recall only the pair of capitalized nouns from each sentence and not the whole sentence itself. They were also encouraged to recall either of the individual nouns in each pair if they could not think of both items. They were allowed 3 min for this free recall. Immediately following this, the subjects attempted to recall the words a second time by using a cued recall. For this, the subjects were given the first nouns of each sentence, randomly ordered on a new sheet of paper, and responded with the corresponding second nouns. The cued recall also lasted for 3 min.

Results

Empirical analysis. All statistical tests were conducted by using the .05 level of significance. Table 1 presents the average vividness ratings for the bizarre and common sentences. For each experiment, mental images for the 12 common sentences were rated by the subjects as significantly more vivid than imagery for the 12 bizarre sentences [Experiment 1: $t(11) = 3.06$, $SD = .22$; Experiment 2: $t(11) = 4.97$, $SD = .19$; Experiment 3: $t(11) = 7.64$, $SD = .12$].

Table 1 also presents the proportion of bizarre and common noun pairs remembered for both free and cued recall. Free-recall performance was measured by the total number of noun pairs in which at least one item was recalled. This is a measure of free recall that has been used

Table 1
Average Vividness Ratings and Proportion of Bizarre and Common Noun Pairs Recalled in Experiments 1, 2, and 3

	Vividness Rating		Free Recall		Cued Recall	
Experiment	Bizarre	Common	Bizarre	Common	Bizarre	Common
1	3.70	4.29	.58	.47	.73	.79
2	3.49	4.42	.59	.54	.70	.71
3	3.30	4.23	.62	.51	.84	.81

Note—Vividness ratings were on a scale of 1 = not vivid through 5 = very vivid.

by previous researchers of bizarre imagery (e.g., McDaniel & Einstein, 1991; PraBaldi et al., 1985) and equals $P(E_1)+P(E_2)+P(E_4)+P(E_5)$. Cued-recall performance was based on the number of noun pairs in which the second item was correctly recalled given the first as a cue, which equals $P(E_1)+P(E_2)+P(E_3)$.

There were no significant differences in cued recall between bizarre and common sentences for any of the three experiments [Experiment 1: $t(32) = 1.56$, $SD = .04$; Experiment 2: $t(31) = 0.28$, $SD = .03$; Experiment 3: $t(29) = 1.00$, $SD = .03$]. However, there was a consistent advantage of bizarre over common sentences for free recall. Although this bizarreness effect was not statistically reliable for Experiment 2 [$t(31) = 1.17$, $SD = .05$], significantly more bizarre noun pairs were freely recalled in Experiment 1 [$t(32) = 2.37$, $SD = .05$] and Experiment 3 [$t(29) = 2.82$, $SD = .04$].

We also examined the order of recall for bizarre and common noun pairs. In a study by Kroll, Schepler, and Angin (1986, Experiment 2), several subjects reported that bizarre imagery came more easily and quickly to mind during recall. Consequently, subjects might remember bizarre images better than common images because they adopt recall strategies that lead them to access bizarre items earlier in the recall sequence. If so, this might have important implications for potential theories of the bizarreness effect.

To conduct this analysis, we performed a median split on each subject's recall protocol, dividing it into equal halves and tabulating the number of bizarre noun pairs recalled within each half. The subjects who recalled one item or less during free recall were not included in this analysis. If subjects adopt recall strategies that favor bizarre stimuli, then we would expect more bizarre noun pairs to be recalled in the first half of subjects' protocols than in the second half. However, the results did not support this hypothesis. Bizarre items were recalled slightly more often in the first half of protocols for each experiment, but none of these differences was statistically reliable [Experiment 1: $t(28) = 0.61$, $SD = .023$; Experiment 2: $t(29) = 0.13$, $SD = .25$; Experiment 3: $t(28) = 1.65$, $SD = .25$]. We performed additional statistical tests, using other methods for measuring output order, but these tests also failed to reveal any significant tendency to recall bizarre items earlier in the recall sequence. These results match those found by Kroll et al. (1986), who also observed a small but nonsignificant advantage in recall order for bizarre imagery.

Model analysis. The empirical analysis reveals that a reliable bizarreness effect occurred in two of the three experiments. However, this analysis in itself does not indicate whether the effect was due to storage or retrieval. To determine this, the free- and cued-recall data for the three experiments were used to tabulate the N_i statistics for the multinomial model, which are presented in Table 2. The value of N equalled 198, 192, and 180 for Experiments 1, 2, and 3, respectively, and based on the Monte Carlo simulations reported earlier, these should be sufficient sample sizes for obtaining accurate parameter estimates.

Table 2
N_i Recall Statistics and Goodness of Fit for the Bizarre and Common Imagery in Experiments 1, 2, and 3

Sentence	N_1	N_2	N_3	N_4	N_5	N_6	$G^2(1)$
			Experiment 1				
Bizarre	93	14	37	0	10	44	1.19
Common	80	5	71	2	7	33	3.19
			Experiment 2				
Bizarre	86	14	36	1	11	44	0.92
Common	83	7	47	2	11	42	1.17
			Experiment 3				
Bizarre	103	2	46	0	7	22	6.86*
Common	80	0	65	3	9	23	21.91*

Note—N_1 = both items freely recalled, correct cued recall; N_2 = one and only one item freely recalled, correct cued recall; N_3 = neither item freely recalled, correct cued recall; N_4 = both items freely recalled, incorrect cued recall; N_5 = one and only one item freely recalled, incorrect cued recall; and N_6 = neither item freely recalled, incorrect cued recall; $G^2(1)$ is the loglikelihood ratio statistic. *$p < .01$.

The first step in evaluating the model is to determine its goodness of fit to the data. As we indicated earlier, this is not possible for Case I of the model because there are no degrees of freedom with which to test its fit. However, Case II of the model can be tested for goodness of fit because the restriction on s_1 and s_2 frees up 1 *df*. As Riefer and Batchelder (1988) describe, one way of assessing goodness of fit for multinomial models is the log likelihood ratio statistic G^2, which is asymptotically distributed as a chi-square variable with 1 *df* for Case II. The computer simulations of the model revealed that G^2 provided a good approximation to the chi-square distribution even for samples as small as $N = 50$. Table 2 presents the values of G^2 for the three experiments, and, as can be seen, the model fits the data from Experiments 1 and 2 well but provides a poor fit for Experiment 3. The poor fit for Experiment 3 is not an insurmountable problem because for this data set we applied Case I, the unrestricted version of the model with s_1 estimated separately from s_2.

Next, we used Equations 3 and 4 to obtain MLEs of the model's parameters for each experiment, which are presented in Table 3. Depending on the model's fit to the data, either the values of s_1 and s_2 (Experiment 3) or the constrained parameter s (Experiments 1 and 2) are given. In general, the values of s, s_1, and s_2 are relatively low, indicating that the subjects remembered most noun pairs together and recalled few of the associates as singletons. Also as expected, the values of r_2 are very high, revealing that storage of a noun pair typically resulted in the successful cued recall for that pair. Of more interest, however, are the values of the storage parameter a and the retrieval parameter r_1. A log likelihood ratio test revealed that subjects retrieved significantly more bizarre than common noun pairs (as measured by r_1) in both Experiment 1 [$G^2(1) = 5.45$] and Experiment 3 [$G^2(1) = 5.34$]. The slight retrieval advantage for bizarre imagery in Experi-

Table 3
Parameter Estimates for Experiments 1, 2, and 3

Sentence	a	r_1	r_2	s_1	s_2	s
			Experiment 1			
Bizarre	.73	.65	1.00	—	—	.23
Common	.81	.51	.98	—	—	.10
			Experiment 2			
Bizarre	.72	.63	.99	—	—	.24
Common	.73	.61	.98	—	—	.17
			Experiment 3			
Bizarre	.84	.68	1.00	.06	.24	—
Common	.84	.55	.96	.00	.30	—

Note—a = probability of storing the association; r_1 = probability of retrieving the association in free recall; r_2 = probability of retrieving the association in cued recall; s_1 = probability of recalling a nonretrieved associate as a singleton; s_2 = probability of recalling a nonassociated item as a singleton; s = probability of recalling a nonretrieved or nonassociated item as a singleton.

ment 2 did not reach statistical significance [$G^2(1)$ = 0.20], which matches the nonsignificant empirical results of Experiment 2. In contrast to the retrieval advantage for bizarre imagery, bizarre and common noun pairs did not significantly differ on their storage (as measured by parameter a) for any of the experiments [$G^2(1)$ = 3.42, 0.09, and 0.01 for Experiments 1, 2, and 3, respectively, where 3.84 is the .05 critical value].

Discussion

The empirical results of the experiments generally matched those observed in previous research. A bizarreness effect was obtained for free recall, although the effect only reached statistical significance for two of the three experiments. Moreover, no bizarreness effect was obtained for cued recall. Although this is consistent with the results of prior studies, it is important to note that free recall always immediately preceded cued recall for each experiment. Presenting free recall before cued recall is a common method for comparing these two types of memory tests (e.g., Hirshman, 1988; Tulving & Pearlstone, 1966) and in fact is the same procedure used by previous researchers to compare the free and cued recall of bizarre imagery (e.g., Hirshman et al., 1989, Experiment 2; Kroll et al., 1986, Experiment 1). However, this leaves open the question as to what effect, if any, the free-recall task had on subjects' cued-recall performance. For example, free recall may have acted as an additional learning trial, thereby increasing performance on the cued-recall tests. If so, then the bizarreness effect might turn out differently if cued recall is tested without a preceding free recall. Experiment 4 was conducted to explore this possibility.

EXPERIMENT 4

Method

The subjects were 34 undergraduate students from the same source as in the previous experiments, and the materials were exactly the same as those used in Experiments 1 and 2. The procedure was the same as well, except that the subjects were not given a free-recall test of their memory. To ensure that the same amount of time elapsed for this experiment between list presentation and cued recall, the subjects received an extra 3 min of the distractor task during the time period devoted to free recall in the previous experiments. Thus, the distractor session totalled 5 min for this experiment, and the subjects were given extra distractors to ensure that they spent the entire time on the task. Cued recall then immediately followed the distractor task.

Results and Discussion

Consistent with the results of Experiments 1-3, there was no bizarreness effect for cued recall. In fact, the subjects recalled a higher proportion of common items (.75) than bizarre items (.67), similar to what was observed in Experiments 1 and 2. However, as in all three prior experiments, this difference was not statistically reliable [$t(33)$ = 1.63, SD = .26]. This supports the conclusion that the lack of a bizarreness effect for cued recall in the previous experiments was not due to any confounding effects of the free-recall task.

GENERAL DISCUSSION

Storage–Retrieval Conclusions

The results of the current experiments seem to mirror previous findings on bizarre imagery, since they provide mixed support for the bizarreness effect. This is not surprising considering the difficulty researchers in the past have had in producing this effect (e.g., Kline & Groninger, 1991; Kroll et al., 1986). Specifically, no advantage for bizarre imagery was observed for cued recall, although two of the three experiments produced a significant bizarreness effect for free recall. Moreover, because bizarre items were not output significantly earlier in the recall sequence than were common items, it seems unlikely that the bizarreness effect can be explained solely in terms of subjects' recall strategies or output preferences. However, despite the mixed empirical results, the analysis of the multinomial model produced an interpretation of the bizarreness effect that was quite consistent. When the bizarreness effect was obtained, the model revealed that the effect was due to retrieval and not to storage factors.

The finding that bizarre imagery is retrieved better than common imagery has implications for a number of current theories of the bizarreness effect. For example, McDaniel and Einstein (1986; 1991; Einstein & McDaniel, 1987) have proposed that bizarre images are more memorable because of their distinctivess within memory, but they also point out that the concept of distinctiveness in itself does not specify which cognitive operations are involved in making these memories distinct. The results of the model's analysis, as well as Einstein et al.'s (1989) finding of superior accessibility for bizarre sentences, seem to provide converging evidence that the effect of distinctiveness is somehow to improve the retrievability of bizarre memory traces. This viewpoint is also consistent with other retrieval-based theories of distinctiveness (see Schmidt, 1991, for a review of theories of distinctiveness).

On the other hand, Hirshman et al. (1989) have proposed that distinctiveness alone is insufficient for explaining the bizarreness effect. Instead, they believe that bizarre imagery elicits a "surprise" response that results in more general contextual cues for its recall. If this idea is correct, then the results of the model suggest that one advantage of these extra cues may be to provide additional retrieval opportunities for recovering bizarre imagery. Hirshman et al. (1989) also speculate that the extra cues associated with bizarre imagery may interfere with the retrieval of common imagery, a hypothesis that is also consistent with our modeling results.

The lack of a significant storage advantage for bizarre imagery also speaks against storage-based theories of the bizarreness effect. For example, the model's results would be difficult to explain with attentional hypotheses (Merry, 1980; Wollen & Cox, 1981) that assume that bizarre imagery receives more cognitive effort or processing than does common imagery, unless one were to conclude that extra processing somehow leads to better retrieval without any benefit to storage. However, our results do not rule out the possibility that storage processes may play an important role in memory for common imagery. Einstein et al. (1989) have proposed that common imagery may lead to better integration or association of the elements within an image. This would help explain why bizarre-imagery studies have sometimes observed a commonness effect (i.e., a recall advantage of common over bizarre imagery). With regard to the multinomial model, parameter a measures the association of items; so if Einstein et al. (1989) are correct, the model would predict that the commonness effect, whenever it occurs, should be accounted for by higher values of parameter a. This is a straightforward prediction that is easily testable by future research.

Additionally, it is important to note that storage factors may also play a role in the bizarreness effect, at least in some circumstances. For example, a common observation in bizarre-imagery research is that subjects report bizarre mental imagery to be less vivid than common mental imagery (Einstein et al., 1989; McDaniel & Einstein, 1991), a finding replicated in our experiments. This may create a disadvantage for bizarre images that could inhibit any potential storage benefits they may have. One way of better controlling for the vividness factor may be to present bizarre and common *pictures* to subjects. Presumably, pictorial stimuli should be equally vivid whether they are bizarre or common, and this may better reveal any potential storage benefits of bizarre imagery. Einstein and McDaniel (1987) have pointed out that experimenter-generated versus subject-generated imagery is a potentially important variable that has yet to be systematically investigated by researchers in this area. Further experiments and modeling along these lines would be helpful in determining what role, if any, storage processes may play in the bizarreness effect.

Advantages of the Model

The model presented in this article provides an analysis of free- and cued-recall data that is similar to, but also extends, previous research in which an operational approach based on empirical statistics was used. As stated in the introduction, memory phenomena that occur for free recall but not for cued recall are typically attributed to retrieval factors (e.g., Tulving & Psotka, 1971), whereas phenomena that occur for both free and cued recall are attributed to storage (e.g., Kail, Hale, Leonard, & Nippold, 1984). The multinomial model presents a model-based rationale for directly measuring these quantities.

There are many situations in which the multinomial model reaches the same conclusions about storage and retrieval as do analyses based on empirical statistics. The current experiments provide an example of this. They reveal that bizarre imagery aids the free recall of items but not their cued recall, an empirical observation that replicates previous findings. This is also the same pattern of results that traditionally has led researchers studying other memory phenomena to conclude that retrieval, and not storage factors, is the cause of those phenomena. Thus, at least for our study, the conclusions based on the multinomial model exactly match those that one would reach using the empirical statistics alone. In fact, it is easy to show that whenever a phenomenon occurs for free but not for cued recall, the model must interpret such a result as arising from retrieval and not from storage.

However, it is important to realize that this is only one possible pattern of results that could occur when comparing free and cued recall. As it turns out, there are many situations in which the model provides a different and much clearer picture of the relative contributions of storage and retrieval than do empirical statistics. For example, one advantage of the modeling approach is in those cases in which a memory phenomenon occurs for both free and cued recall. This pattern of results is usually interpreted by the empirical approach as support for a storage explanation of the phenomenon. However, this conclusion leaves unanswered whether storage alone is the cause or whether retrieval factors may also play a role. This can potentially confuse the issue if one is forced to rely solely on empirical statistics for the answer.

A good illustration of this problem can be seen in an experiment by Kail et al. (1984). They tested the memory abilities of language-impaired children on a number of different measures, including free and cued recall. They reasoned that if storage hypotheses of language impairment are correct, then memory deficits exhibited by these children should occur for both free and cued recall. In addition, if language-impaired children are also deficient in retrieval skills, then their memory deficits should be greater for free recall than for cued recall. The empirical results showed that language-impaired children were in fact poorer than normal children on both free and cued recall, and from this, Kail et al. (1984) were able to confirm the storage

hypothesis. But a nonsignificant interaction indicated that this deficit was equally strong for each type of recall, from which they concluded that "the situation regarding the retrieval hypothesis [was] more complicated" (p. 46). The problem that they encountered was that although the analysis of free and cued recall did not support retrieval explanations, other measures did. Kail et al. even included the analysis of their own multinomial model for repeated free recall, which pointed to both storage and retrieval factors. They eventually concluded that language impairment affects both storage and retrieval, although this conclusion was quite different from the one implied solely by the comparison of free and cued recall.

In fact, the limitations of relying only on an empirical comparison of free and cued recall can now be made even clearer with the following demonstration. Suppose two groups of subjects differ in their memory abilities; also suppose that they differ on both storage and retrieval processes. To quantify this, assume that Case II of the model is correct, and that the true parameter values for Group 1 are $a = .7$, $r_1 = .7$, $r_2 = .9$, $s = .1$ and the true values for Group 2 are $a = .3$, $r_1 = .3$, $r_2 = .9$, $s = .1$. As can be seen, these values have been selected so that the groups differ equally on storage and retrieval and do not differ for parameters r_2 and s. If these two groups were tested on their free and cued recall in a hypothetical experiment, Equation 1 could be used to compute the resultant E_i probabilities, which in turn could be used to compute the probabilities of free and cued recall for each group. The results would yield a free-recall probability of .54 and a cued-recall probability of .63 for Group 1, with corresponding values of .18 and .27 for Group 2. A close look at these values shows that the empirical difference between the two groups is precisely the same for both free and cued recall (a difference of .36 in each case). If we adopt Kail et al.'s (1984) earlier reasoning, we would conclude from this result that the two groups differ on storage—a correct conclusion. But we would also be forced to conclude that the empirical results do not support a retrieval explanation for the differences, despite the fact that the example was engineered so that retrieval effects were equally as strong as storage effects.

Of course, one reason for the failure to detect retrieval differences in the above example comes from the fact that free and cued recall were the only measures examined. One could argue that a better approach would be either to conduct a more comprehensive analysis of the data or to examine a larger number of different empirical measures of storage and retrieval similar to what was done by Kail et al. (1984). Unfortunately, researchers often use just one or two dependent measures to examine memory processes, so we would agree that the use of multiple measures of memory is more advantageous for studying underlying memory processes. However, the above examples reinforce our earlier point that empirical approaches alone may be limited in their ability to reveal the relative contributions of storage and retrieval. We would advocate that researchers who wish to supplement their data analysis should, if possible, conduct some type of modeling analysis in addition to a comprehensive empirical analysis. This could be done by developing a new model specifically for the study or by using a preexisting model such as the one developed here.

The above examples also illustrate the advantage of the multinomial model for measuring the relative effects of storage and retrieval when both processes contribute to a memory phenomenon. Clearly, storage and retrieval are important factors to some degree in most studies of memory. The measurement of these quantities in the form of the model's parameters can give researchers a more precise, quantitative measure of these processes. Riefer and Batchelder (1988) even show how one can derive confidence intervals and confidence regions for the parameter estimates. This is especially useful for researchers whose theories are concerned with the relative contributions of storage and retrieval. The model is capable of determining not only when storage or retrieval make significant contributions to memory but can also help to establish the relative strength of each process.

A quick analysis of the model also shows that it has an interesting relation to traditional statistics when there is a small number of E_4 events (i.e., when both items of an associate are freely recalled but cued recall fails). As the number of E_4 events approaches zero, it is easy to see from Equation 3a that the storage parameter a is estimated by the expression $(N_1+N_2+N_3)/N$. This is just the proportion of correct cued recall responses, so in this case the model's analysis is identical to the conventional method of assuming that cued recall measures storage factors. One implication of this is that, if N_4 is close to zero, there can be no storage differences in parameter a unless there are cued-recall differences. Typically, E_4 events are relatively rare, but they do occur. Moreover, the possibility exists for data sets when there is a substantial number of E_4 events (cf. Thomson & Tulving, 1970; Tulving & Wiseman, 1975). Unfortunately, the traditional assumption that cued recall measures storage leads to an unappealing interpretation of these E_4 events, since it assumes that the associate has not been stored (incorrect cued recall) but somehow has been retrieved (correct free recall). In contrast, the multinomial model does a more satisfactory job of handling the E_4 events. An E_4 event is interpreted by the model as a series of cognitive processes in which the associate is stored, is retrieved for free recall, but then fails to be retrieved for cued recall. Researchers may then hypothesize about this failure. In addition, as the number of E_4 events increases, the storage-parameter estimate also increases, whereas the main retrieval parameter r_1 stays constant. This interpretation seems more psychologically reasonable than that afforded by the conventional method.

As researchers conduct further empirical experiments to explore the theoretical basis behind the bizarreness effect, the role of different cognitive processes will continue to be an important focus for theorists in this area. The general point of this article has been to show how

the multinomial model presented here can provide a useful measurement tool in aiding this research. Moreover, it should also be pointed out that the model is capable of exploring many more issues in the area of human memory than just the bizarreness effect, especially storage and retrieval issues that can be examined by using the paired-associate paradigm. For example, the model is capable of exploring the storage and retrieval basis of other mnemonic techniques, such as the keyword system (Desrochers & Begg, 1987), and can be used to compare memory processes in different subject populations or for different types of stimuli (Riefer & Batchelder, 1991a). With the development of the multinomial model presented here, theorists now have a number of storage-retrieval models that can be applied to different learning paradigms, including Batchelder and Riefer's (1980, 1986) multinomial model for category pairs in free recall, Chechile's (1987) storage-retrieval model that examines cued recall and recognition of paired associates, and Kail et al.'s (1984) multinomial model for repeated free recall. Research concerning storage and retrieval factors in memory would benefit from the continued use of these or similar models to supplement traditional empirical analyses of memory phenomena.

REFERENCES

ANDREOFF, G. R., & YARMEY, A. D. (1976). Bizarre imagery and associative learning: A confirmation. *Perceptual & Motor Skills*, **43**, 143-148.

BATCHELDER, W. H., & RIEFER, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, **87**, 375-397.

BATCHELDER, W. H., & RIEFER, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical & Statistical Psychology*, **39**, 129-149.

BERGFELD, V. A., CHOATE, L. S., & KROLL, N. E. A. (1982). The effect of bizarre imagery on memory as a function of delay: Reconfirmation of interaction effect. *Journal of Mental Imagery*, **6**, 141-158.

CHECHILE, R. A. (1987). Trace susceptibility theory. *Journal of Experimental Psychology: General*, **116**, 203-222.

CORNOLDI, C., CAVEDON, A., DE BENI, R., & PRA BALDI, A. (1988). The influence of the nature of material and of mental operations on the occurrence of the bizarreness effect. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **40**, 73-85.

COX, S. D., & WOLLEN, K. A. (1981). Bizarreness and recall. *Bulletin of the Psychonomic Society*, **18**, 244-245.

CRAIK, F. I. M., & MCDOWD, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 474-479.

DESROCHERS, A., & BEGG, I. (1987). A theoretical account of encoding and retrieval processes in the use of imagery-based mnemonic techniques: The special case of the keyword method. In M. A. McDaniel & M. Pressley (Eds.), *Imagery and related mnemonic processes: Theories, individual differences, and applications* (pp. 56-77). New York: Springer-Verlag.

EINSTEIN, G. O., & MCDANIEL, M. A. (1987). Distinctiveness and the mnemonic benefits of bizarre imagery. In M. A. McDaniel & M. Pressley (Eds.), *Imagery and related mnemonic processes: Theories, individual differences, and applications* (pp. 78-102). New York: Springer-Verlag.

EINSTEIN, G. O., MCDANIEL, M. A., & LACKEY, S. (1989). Bizarre imagery, interference, and distinctiveness. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 137-146.

EMMERICH, H., & ACKERMAN, B. (1979). A test of bizarre interaction as a factor in children's memory. *Journal of Genetic Psychology*, **134**, 225-232.

HANLEY, J., & MORRIS, P. (1987). The effects of amount of processing on recall and recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **39**, 431-449.

HIRSHMAN, E. (1988). The expectation-violation effect: Paradoxical effects of semantic relatedness. *Journal of Memory & Language*, **27**, 40-58.

HIRSHMAN, E., WHELLEY, M. M., & PALIJ, M. (1989). An investigation of paradoxical memory effects. *Journal of Memory & Language*, **28**, 594-609.

HOGAN, R. M., & KINTSCH, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior*, **10**, 562-567.

HOGG, R. V., & CRAIG, A. T. (1978). *Introduction to mathematical statistics*. (3rd ed.). New York: Macmillan.

IACCINO, J. F., DVORAK, E., & COLER, M. (1989). Effects of bizarre imagery on the long-term retention of paired associates embedded within variable contexts. *Bulletin of the Psychonomic Society*, **27**, 114-116.

IACCINO, J. F., & SOWA, S. J. (1989). Bizarre imagery in paired-associate learning: An effective mnemonic aid with mixed context, delayed testing, and self-paced conditions. *Perceptual & Motor Skills*, **68**, 307-316.

KAIL, R., HALE, C. A., LEONARD, L. B., & NIPPOLD, M. A. (1984). Lexical storage and retrieval in language-impaired children. *Applied Psycholinguistics*, **5**, 37-49.

KLINE, S., & GRONINGER, L. D. (1991). The imagery bizarreness effect as a function of sentence complexity and presentation time. *Bulletin of the Psychonomic Society*, **29**, 25-27.

KROLL, N. E. A., SCHEPLER, E. M., & ANGIN, K. T. (1986). Bizarre imagery: The misremembered mnemonic. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 42-53.

KROLL, N. E. A., & TU, S.-F. (1988) The bizarre mnemonic. *Psychological Research*, **50**, 28-37.

LORAYNE, H., & LUCAS, J. (1974). *The memory book*. New York: Stein & Day.

MCDANIEL, M. A., & EINSTEIN, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 54-65.

MCDANIEL, M. A., & EINSTEIN, G. O. (1991). Bizarre imagery: Mnemonic benefits and theoretical implications. In R. H. Logie & M. Denis (Eds.), *Mental images in human cognition* (pp. 183-192). New York: Elsevier Science Publishers.

MCNULTY, J. A., & CAIRD, W. (1966). Memory loss with age: Retrieval or storage? *Psychological Reports*, **19**, 229-230.

MERRY, R. (1980). Image bizarreness in incidental learning. *Psychological Reports*, **46**, 427-430.

O'BRIEN, E. J., & WOLFORD, C. R. (1982). Effect of delay in testing on retention of plausible versus bizarre mental images. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 148-152.

PRA BALDI, A., DE BENI, R., CORNOLDI, C., & CAVEDON, A. (1985). Some conditions of the occurrence of the bizarreness effect in free recall. *British Journal of Psychology*, **76**, 427-436.

RIEFER, D. M., & BATCHELDER, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, **95**, 318-339.

RIEFER, D. M., & BATCHELDER, W. H. (1991a). Age differences in storage and retrieval: A multinomial modeling analysis. *Bulletin of the Psychonomic Society*, **29**, 415-418.

RIEFER, D. M., & BATCHELDER, W. H. (1991b). Statistical inference for multinomial tree models. In J.-C. Falmagne & J.-P. Doignon (Eds.), *Mathematical psychology: Current developments* (pp. 313-336). Berlin: Springer-Verlag.

SCHMIDT, S. R. (1991). Can we have a distinctive theory of memory? *Memory & Cognition*, **19**, 523-542.

SCHONFIELD, D., & ROBERTSON, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology*, **20**, 228-236.

SMITH, A. D. (1980). Age differences in encoding, storage, and retrieval. In L. Poon, J. Fozard, L. Cermak, D. Arenberg, & L. Thompson (Eds.), *New directions in memory and aging* (pp. 23-46). Hillsdale, NJ: Erlbaum.

THOMSON, D. M., & TULVING, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology*, **86**, 255-262.

TULVING, E., & PEARLSTONE, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior*, **5**, 381-391.

TULVING, E., & PSOTKA, J. (1971). Retroactive inhibition in free recall: Inaccessibility of information available in the memory store. *Journal of Experimental Psychology*, **97**, 1-8.

TULVING, E., & WISEMAN, S. (1975). Relation between recognition and recognition failure of recallable words. *Bulletin of the Psychonomic Society*, **6**, 79-82.

WEBBER, S. M., & MARSHALL, P. H. (1978). Bizarreness effects in imagery as a function of processing level and delay. *Journal of Mental Imagery*, **2**, 291-300.

WOLLEN, S. B., & COX, S. D. (1981). Sentence cuing and the effectiveness of bizarre imagery. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 386-392.

WOLLEN, S. B., WEBER, A., & LOWRY, D. H. (1972). Bizarreness versus interaction of mental images as determinants of learning. *Cognitive Psychology*, **3**, 518-523.

APPENDIX

Sentences Used in Experiments 1, 2, and 3

GIRL-DOLL
Common: The GIRL kissed the DOLL.
Bizarre: The GIRL boiled the DOLL.

MAID-AMMONIA
Common: The MAID spilled the AMMONIA.
Bizarre: The MAID drank the AMMONIA.

CHEF-PICKLE
Common: The CHEF sliced the PICKLE.
Bizarre: The CHEF smoked the PICKLE.

DOG-BICYCLE
Common: The DOG chased the BICYCLE.
Bizarre: The DOG rode the BICYCLE.

LAWYER-CHAIR
Common: The LAYWER sat on the CHAIR.
Bizarre: The LAWYER argued with the CHAIR.

GOLDFISH-BOWL
Common: The GOLDFISH was swimming in the BOWL.
Bizarre: The GOLDFISH was eating out of the BOWL.

COCKROACH-STOVE
Common: The COCKROACH appeared on the STOVE.
Bizarre: The COCKROACH moved the STOVE.

SNOWFLAKE-MOUNTAIN
Common: The SNOWFLAKE fell on the MOUNTAIN.
Bizarre: The SNOWFLAKE climbed the MOUNTAIN.

Additional Sentences Used in Experiments 1 and 2

STUDENT-SANDWICH
Common: The STUDENT made the SANDWICH.
Bizarre: The STUDENT stabbed the SANDWICH.

MINISTER-BIBLE
Common: The MINISTER read the BIBLE.
Bizarre: The MINISTER ate the BIBLE.

BANKER-NEWSPAPER
Common: The BANKER folded the NEWSPAPER.
Bizarre: The BANKER floated on the NEWSPAPER.

SHOES-MILK
Common: The SHOES were placed by the MILK.
Bizarre: The SHOES were filled with MILK.

Additional Sentences Used in Experiment 3

CAR-FENCE
Common: The CAR drove past the FENCE.
Bizarre: The CAR pets the FENCE.

PRIEST-BIBLE
Common: The PRIEST read the BIBLE.
Bizarre: The PRIEST ate the BIBLE.

PLANT-TELEVISION
Common: The PLANT rested on top of the TELEVISION.
Bizarre: The PLANT screamed at the TELEVISION.

LAMP-BOOK
Common: The LAMP shined on the BOOK.
Bizarre: The LAMP read the BOOK.

(Manuscript received August 10, 1990;
revision accepted for publication January 15, 1992.)