Bulletin of the Psychonomic Society
1989, 27 (3), 215-218

The bizarre mnemonic: The effect of retention interval and mode of presentation

CARRIE L. ZOLLER, JEFF S. WORKMAN, and NEAL E. A. KROLL
University of California, Davis, Davis, California

Two experiments are reported, in which subjects were told to use visual images to help them remember lists of words. Both experiments yielded the typical results with immediate free recall: In mixed lists, words in bizarre contexts are remembered better than those in common contexts, but no such advantage exists for pure lists. In Experiment 1, memory was also tested after 48 h, and no evidence was found to suggest that the advantage for bizarre contexts continued to increase during this longer retention interval, or that it is even present for groups tested only after this delay. That is, the bizarre context effect (BCE) appears to be limited to brief retention intervals. In Experiment 2, words presented with pictures were found to be remembered better than those presented with sentences, but there was no interaction of this picture advantage with the BCE. In Experiment 1 no correlation between extroversion and BCE magnitude was found, and Experiment 2 yielded no correlation between mental image ability and BCE magnitude. However, subjects with high mental-image ability scores remembered more from pure bizarre lists than did those with low mental-image scores.

Professional mnemonists have long advocated using bizarre imagery to improve memory (for a sympathetic review, see Yarmey, 1984). In addition, over the years, a number of experimental psychologists have advanced reasons for why the bizarreness of the imagery should aid memory (e.g., McDaniel & Einstein, 1986). Others (e.g., Kroll & Tu, 1988) have argued that a bizarre context does not, in general, improve memory, and that in the specific cases in which a bizarre context does have a positive effect on memory, it is not at all clear that this bizarre context effect (BCE) necessarily involves the use of imagery.

In an attempt to salvage an imagery interpretation of the BCE, Wollen and Margres (1987) listed "four key phenomena that any theory of bizarre imagery must accommodate": (1) Bizarreness is most likely to have a beneficial effect on sentence free recall (FR), a negative effect on the average number of words recalled per sentence, and a zero or negative effect on cued recall (CR) and recognition. (2) The positive BCE is restricted to mixed lists. (3) More time is required to form bizarre images. (4) Bizarre images are rated as being less vivid than common images.

In order to "accommodate" all of these typically found phenomena, Wollen and Margres developed the imagery multiprocess (IMP) model. According to IMP, a person forming common images relies upon preexisting schemata that result in fast, well-formed, vivid images. To form a bizarre image, however, a new image must be created. This requires time, and the image is less likley to be well formed and vivid. The uniqueness of the bizarre image will make it more distinctive, and hence, in a mixed list, it will be easier to find in memory than a common image; but it will likely be less integrated, so that finding part of such an image will not necessarily ensure the recovery of all of it.

Although the IMP model appears to handle most of the existing BCE data better than its competitors, there are findings that this model does not explain well. For example, bizarre contexts seem to improve memory only when there are relatively few of them in a list (Kroll & Tu, 1988, Experiments 4 and 5). Also, although the bizarre items in a mixed list may be learned faster than the common items, the rate of forgetting appears to be relatively constant. That is, a greater proportion of bizarre items is recalled with immediate FR, but that difference remains relatively constant after longer retention intervals (Kroll & Tu, 1988, Experiment 6). One purpose of our Experiment 1 was to replicate this latter finding and to include a comparison group tested only at the longer retention interval.

Another disturbing fact is that not only the IMP model, but the very data upon which it is based, does not fit the claim from professional mnemonists that bizarre imagery should actually improve memory—not simply determine which items in a mixed list will be more likely to be recalled, usually at the expense of other items so that the overall number of items recalled from the list will not increase. In Experiment 1, we assessed the possibility that the difference between the claims of mnemonists and findings from college students might be partially explained by personality factors. That is, if, as seems likely, professional mnemonists are more likely to be more extroverted than most college students, then one might expect to find a correlation between a person's extroversion score and that person's BCE.

Requests for reprints should be sent to Neal Kroll, Department of Psychology, University of California, Davis, CA 95616.

EXPERIMENT 1

Method

Procedure and Design. The subjects attended two sessions, 48 h apart. The first session (Day 1) began with the subjects' being instructed to close their eyes and to attend closely to a selection of recorded music. The selection was immediately preceded by a male-voiced "now" and followed by "end."

The subjects then saw a series of 18 slides, each containing three to-be-remembered (TBR) words embedded in a high-imagery sentence, and they were asked to mentally form the images suggested by the sentences, in order to help themselves remember the triplets for a later test. The TBR triplets were identical in all conditions, but the suggested imagery varied across context (bizarre or common) and design (pure or mixed) conditions. Some subjects saw only sentences suggesting bizarre images (pure bizarre) or only those suggesting common images (pure common). Others saw a mixture of nine bizarre and nine common sentences. The contexts were sequenced randomly in mixed conditions, with the restriction that there be no more than three of the same context type successively. Regardless of context-design condition, the subjects were told that the mnemonic they were to use was the most effective. For example, mixed-list subjects heard that research had shown that list memory is best with a mixture of bizarre and common imagery. The experimenter said "now" at the onset of the first slide and "end" as the last slide left the screen.

In addition to the context-design differentiation, subjective time estimates (STE) were made during either Day 1 or Day 2, and the subjects received either one or two FR tests.

The subjects in two-test conditions were given an FR test immediately after the presentation of the slides. Six minutes were allowed for FR; then they rated both the complexity and their enjoyment of the music. Half of these subjects then completed an STE, comparing the music duration with the total time the slides were on the screen. At the end of Day 1, all the subjects were reminded to return at the same time in 2 days when they would be asked to do something different. Day 2 began with a surprise FR test for all subjects. The subjects who had not completed the STE did so after this test.

The subjects in the one-test conditions were not given an FR test on Day 1, only music-opinion questions and the STE. They were then told that the experimenter had misled them about the memory test in order to obtain the STE. They too received a surprise FR test at the beginning of Day 2.

All subjects ended the experiment by completing a personality inventory that was scored for extroversion.

Subjects. The subjects were undergraduates at the University of California, Davis who received extra credit in an introductory psychology or rhetoric class in return for their participation. They were tested in groups that varied in size from 1 to 12.

A total of 144 subjects participated in the two-test conditions, with 24 subjects in each of the 3×2 orthogonal conditions formed by the context-design (pure bizarre, pure common, or mixed) and STE (Day 1 or Day 2) conditions. An additional 144 subjects were in the one-test (Day 2) conditions, with 48 subjects in each of the three context-design conditions.

Materials. The music was a selection from "Heterophonie" by Mauricio Kagel; it had been chosen for its lack of constant rhythm and its apparent disorganization.

Fifteen sentences were taken from McDaniel and Einstein's (1986) list and 3 from Kroll and Tu (1988). The TBR words were in underlined capital letters. The presentation of the slides was controlled by an electronic timer, which presented each slide at a constant rate of 10±.10 sec, for a total time equal to that of the musical selection.

The more direct extroversion measure was obtained from the Eysenck personality inventory, Form A (Eysenck, 1970). A more indirect measure of extroversion (Lomranz, 1983) was obtained from the STE task. The subjects were presented with two lines: a medium-length top line representing the time of the music, and a longer, calibrated bottom line on which they marked how long they thought the slides had lasted in comparison with the music.

Results

There were no consistent patterns of correlations resulting from the personality inventory nor from the STE. Consequently, these measures will be ignored in the presentation of the results. The FR results are presented in Table 1. The analysis procedure (Erlebacher, 1977, 1978) for contrasting between- and within-subjects effects was applied to both sentence and word FR data and the patterns of significant results from the two data sets were almost identical.[1] The results of the slightly more sensitive sentence FR analysis will be reported.

Subjects tested both days. There was, not surprisingly, a strong effect of retention interval [$F(1,45) = 328.59$, $MS_e = 48.00$, $p < .001$]. The Day 1 FR pattern shows the by now typical finding of an advantage for bizarre contexts only in the mixed condition. The Day 2 pattern is, however, different. There is still an advantage for bizarre contexts in the mixed condition, but now there is an advantage for common contexts in the pure conditions. These patterns resulted in the following significant interactions: retention interval × context [$F(1,43) = 6.99$, $MS_e = 51.62$, $p < .025$], retention interval × design type (pure or mixed) [$F(1,45) = 4.70$, $MS_e = 48.00$, $p < .05$], and context × design type [$F(1,53) = 5.846$, $MS_e = 281.14$, $p < .025$].

Thus, it would appear that the advantage from bizarre contexts in mixed lists continues over time, but that an advantage from common over bizarre contexts develops over time in these pure lists. This is the same result pattern found by Kroll and Tu (1988) with lists of 12 triplets rather than the 18 used here.

Subjects tested only on Day 2. These subjects provide a baseline for an estimate of the degree to which FR performance on Day 2 was influenced by retrieval activities during Day 1. No significant differences were found among the conditions with these subjects (all three *F*s—context, design type, and interaction—were less than 1.0). Thus, the advantage for bizarre contexts within mixed lists on Day 2 with twice-tested subjects may simply reflect the added practice that they received with the bizarre-context items during the first FR test. However, it is less easy to explain the pure-list differences between the Day 2 performances of the two-test and Day 2 only test subjects.

Discussion

The immediate FR results confirmed earlier findings and the predictions of the IMP model. There was a mixed-list BCE, but no such ef-

Table 1
Percentage Correct by Free Recall as a Function of Context-Design and Retention Interval (Scores of Subjects Tested only on Day 2 are in Parentheses)

	Sentences		Words	
Context-Design	Day 1	Day 2	Day 1	Day 2
Bizarre pure	49.7	36.3 (23.5)	42.0	27.5 (14.0)
Common pure	49.7	40.4 (22.7)	44.0	33.1 (14.3)
Bizarre mixed	56.3	40.0 (21.5)	48.1	32.3 (13.3)
Common mixed	48.1	35.6 (22.9)	41.3	27.9 (13.0)

fect with pure lists. In addition, the overall average of items recalled from the mixed list was almost identical to that recalled in the pure common list. The second (delayed) FR test did not support the notion that bizarre items were forgotten at a slower rate. In fact, it appeared that the bizarre items were forgotten faster: the mixed-list BCE shrank over time and a small reversed BCE began to appear with the pure lists.

The delayed-only FR test did not result in any differences among the various conditions. This would seem to suggest that the BCE is the result of a very short-term process—contrary to the arguments of those who contend that bizarre imagery should result in longer-lasting memories. The brevity of the BCE may fit well with IMP in that the "newly created" bizarre images might be expected to be less stable than the common images, which are based upon preexisting schemata.

Given our initial hopes based on findings from a pilot experiment, the lack of any correlations among the personality, subjective time, and FR measures was quite disappointing. However, if the BCE does in fact rely upon mental imagery, one would expect some correlations between BCE and individual differences in mental image ability. In Experiment 2, we attempted to assess this possibility.

EXPERIMENT 2

Earlier attempts to find a relationship between a BCE and mental-image ability as indicated by self-evaluation measures have not been successful (e.g., Bergfeld, Choate, & Kroll, 1982). In the present experiment, therefore, we assessed mental-image ability with the more direct clock-imagery task (Paivio, 1978).

The degree to which a BCE relies upon mental visual imagery should also be related to the type of stimulus. That is, if it is the image per se—rather than, say, some semantic or comic aspect—of the bizarre stimulus, then presenting the stimuli by means of pictures rather than sentences should result in a stronger BCE. At least, one might expect this for those subjects with less visual image ability who would have difficulty forming "pictures in their heads" on the basis of the sentences. Although many experimenters have used picture stimuli before, our purpose was to present the same word lists by means of both pictures and sentences, and, at the same time, to search for individual differences with regard to mental visual-image ability.

Method

Procedure and Design. The subjects were first allowed 6 min to complete a speed-test version of Paivio's (1978) clock-imagery task, which was used to measure subjects' image abilities. The test sheet provided a calibrated circle and 45 pairs of clock times written digitally. The subjects were asked to imagine each pair as if the times were being shown on two clocks, to compare these two images, and to indicate for which time the minute and the hour hands formed the smaller angle. For example, the correct answer for the pair 4:15 and 9:05 was 4:15.

The subjects then saw a series of 24 slides, each containing two words that were to be remembered for a later memory test. In the sentence conditions, underlined capitalized word pairs were embedded in sentences describing simple scenes. In the picture conditions, the word pairs were presented alongside a cartoon sketch of the scene described by the sentence for that pair. For example, one word pair was "SPIDER-FLOWER." The bizarre sentence for this pair was "The SPIDER sniffs the FLOWER," and the bizarre picture was a cartoon sketch of a spider holding a flower to its nose. The common sentence for this pair was "The SPIDER spins a web on the FLOWER," and the common picture was a sketch of a spider in a web attached to a flower. The subjects in the sentence conditions were instructed to form mental images of the scenes suggested by the scenes. All subjects were instructed to use the scenes to help themselves remember the word pairs.

Both the sentence and picture conditions were divided into three context-design conditions: pure bizarre (all sentences or pictures were of bizarre scenes); pure common (all scenes were common); and mixed. The mixed condition was further subdivided into the 6/18 condition, with a one-third/two-thirds division of bizarre and common scenes; the 12/12 condition, with one-half/one-half division, and the 18/6 condition, with a two-thirds/one-third division. In addition, there were two different sequences of word pairs in each of the conditions, and in the mixed conditions, any given word pair was seen by approximately half of the subjects with a bizarre and half with a common context. As in Experiment 1, all subjects were told that the mnemonic they were using was believed to be the most effective.

Upon completion of the slides, the subjects received an FR test of the word pairs, followed by a CR test in which one word of each pair was used as a retrieval cue for the second.

Subjects. The subjects were 300 undergraduate students at the University of California, Davis, who received extra credit in an introductory psychology class in return for their participation. There were 25 subjects in each of the modality (sentences or pictures) × context-design (pure bizarre, pure common, mixed 6/18, or mixed 18/6) combinations. There were 50 subjects in each of the two mixed 12/12 conditions. The subjects were tested in groups of 12 or less.

Results

The results of both the FR and the CR tests are presented in Table 2. Separate analyses were performed on the FR and CR data sets.

The first set of analyses to be reported used the Erlebacher procedure for contrasting between- and within-subjects effects. These analyses compared the pure lists with the mixed 12/12 lists. The sums and sums of squares from mixed 12/12 conditions were divided by two, so that the analysis weighed these conditions as if they too contained 25 subjects each. This had a slightly conservative effect on the analyses.

The subjects who viewed pictures recalled significantly more words than did those who viewed sentences [FR, $F(1,92) = 16.56$, $MS_e = 168.84$, $p < .001$; CR, $F(1,87) = 5.22$, $MS_e = 340.01$, $p < .05$]. In addition, there was a significant context × design interaction [FR, $F(1,96) = 8.12$, $MS_e = 157.85$, $p < .01$; CR, $F(1,138) = 7.12$, $MS_e = 172.56$, $p < .01$]. This interaction represents an advantage for common contexts with pure lists and an advantage for bizarre contexts with mixed lists with both

Table 2
Percentage of Words Correctly Recalled as a Function of Context-Design and Modality

	Modality					
	Sentences			Pictures		
Context-Design	Bizarre	Common	Total	Bizarre	Common	Total
		Free Recall				
Bizarre pure	44.1		44.1	51.8		51.8
Mixed 6/18	46.3	36.5	39.8	58.5	44.5	49.2
Mixed 12/12	43.1	39.3	41.2	52.4	44.5	48.5
Mixed 18/6	41.8	38.0	40.5	50.3	45.0	48.5
Common pure		48.3	48.3		56.1	56.1
		Cued Recall				
Bizarre pure	75.3		75.3	81.0		81.0
Mixed 6/18	84.0	83.5	83.7	87.0	82.0	83.7
Mixed 12/12	79.3	77.3	78.3	85.3	83.5	84.4
Mixed 18/6	76.5	79.0	77.3	88.3	85.0	87.2
Common pure		83.2	83.2		89.2	89.2

picture and sentence presentations. Not surprisingly, the common advantage with pure lists is greater for CR, while the bizarre advantage with mixed lists is greater for FR. For FR, but not for CR, there was also an effect of design, with more words' being remembered from pure lists than from mixed lists [FR, $F(1,92) = 8.10$, $MS_e = 168.84$, $p < .01$; CR, $F(1,87) < 1.0$].

Additional analyses were performed on the total number of words recalled from the lists, regardless of their context (the Total columns in Table 2). In these modality (pictures or sentences) × context-design (pure common, pure bizarre, mixed 6/18, mixed 12/12, and mixed 18/6) analyses, the data from the mixed 12/12 was again treated as though it represented only 25 subjects in each modality group. Pictures, of course, again resulted in better recall than sentences [FR, $F(1,240) = 39.97$, $MS_e = 23.15$, $p < .001$; CR, $F(1,240) = 10.24$, $MS_e = 10.76$, $p < .01$].

The context-design dimension of the analysis was broken into the following contrasts: pure versus mixed, pure bizarre versus pure common, and residual. Pure common resulted in better overall recall than pure bizarre [FR, $F(1,240) = 4.58$, $p < .05$; CR, $F(1,240) = 8.56$, $p < .01$]. In addition, for FR, but not for CR, there was a significant advantage of the pure design over the mixed [FR, $F(1,240) = 17.74$, $p < .01$; CR, $F(1,240) < 1.0$]. There was no indication of any interaction with modality.

Correlations between clock-imagery ability and total FR performance were significant in both the picture and sentence pure bizarre conditions [$r(23) = .383$ and $.486$]. These correlations were much smaller for the pure common conditions (.029 and .035) and were generally much smaller for the mixed conditions (ranging from −.179 to .020) except for the picture mixed 6/18 condition (.358). In the mixed conditions, correlations were also found between clock-imagery scores and bizarre common difference scores, and between clock-imagery scores and bizarre FR scores. No consistent patterns were observed.

The pattern of correlations, then, suggested that, at best, the clock-imagery test could only offer moderate predictability for the pure bizarre FR performance, and virtually no predictability for FR in the other conditions.

Discussion

The memory advantage of pictures over sentences does suggest that imagery should help memory—but also that imagery generated from sentences does not result in as good a memory as does imagery directly presented as pictures. The absence of a BCE × modality interaction, however, means that there is no support for the idea that a larger BCE would be found with people having more mental visual-image ability (presumably, professional mnemonists). The lack of any pattern of correlations in the mixed conditions with the clock-imagery tests also fails to support any relationship between visual imagery and the BCE. However, correlations were found between clock-imagery scores and memory in the conditions with pure bizarre lists, which suggested that subjects with good mental imagery were better able to deal with the pure bizarre lists—regardless of the mode of presentation.

When there were three TBR words per sentence, Kroll and Tu (1988) had found that the BCE disappeared in lists with as many as 12 bizarre sentences. The use of two TBR words per item (sentence or picture) did appear to extend the BCE to mixed lists with as many as 18 bizarre items. However, the mixed-list conditions resulted in fewer items' overall being remembered than in the pure common list conditions, so the advantage of bizarre imagery for improvement of memory remains to be demonstrated.

REFERENCES

Bergfeld, V., Choate, L., & Kroll, N. (1982). The effect of bizarre imagery on memory as a function of delay: Reconfirmation of the interaction effect. *Journal of Mental Imagery*, **6**, 141-158.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variables. *Psychological Bulletin*, **84**, 212-219.

Erlebacher, A. (1978). The analysis of multifactor experiments designed to contrast the within- and between-subjects manipulation of the independent variables. *Behavior Research Methods & Instrumentation*, **10**, 833-840.

Eysenck, H. J. (1970). *The biological basis of personality*. Springfield, IL: Thomas.

Kroll, N., & Tu, S. (1988). The bizarre mnemonic. *Psychological Research*, **50**, 28-37.

Lomranz, J. (1983). Time estimation as a function of stimulus complexity & personality. *Social Behavior & Personality*, **11**, 77-82.

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

Paivio, A. (1978). Comparisons of mental clocks. *Journal of Experimental Psychology: Human Perception & Performance*, **4**, 61-71.

Wollen, K., & Margres, M. (1987). Bizarreness and the multi-process model. In M. McDaniel & M. Pressley (Eds.), *Imaginal and mnemonic processes* (pp. 103-128). New York: Springer-Verlag.

Yarmey, A. (1984). Bizarreness effects in mental imagery. In A. Sheikh (Ed.), *International review of mental imagery* (Vol. 1, pp. 57-76). New York: Human Sciences Press.

NOTE

[1] Word FR simply measures the percentage of the total words recalled from the various TBR word triplets. In sentence FR, a sentence is scored as recalled if at least one word from the TBR triplet from that sentence is remembered. Sentence FR is typically found to be the most sensitive to the bizarre context advantage in mixed lists (e.g., by McDaniel & Einstein, 1986).

(Manuscript received October 27, 1988.)