Journal of Experimental Psychology:
Human Learning and Memory
1978, Vol. 4, No. 6, 592-604

The Generation Effect: Delineation of a Phenomenon

Norman J. Slamecka and Peter Graf
University of Toronto, Toronto, Canada

Five experiments are reported comparing memory for words that were generated by the subjects themselves with the same words when they were simply presented to be read. In all cases, performance in the generate condition was superior to that in the read condition. This held for measures of cued and uncued recognition, free and cued recall, and confidence ratings. The phenomenon persisted across variations in encoding rules, timed or self-paced presentation, presence or absence of test information, and between- or within-subjects designs. The effect was specific to the response items under recognition testing but not under cued recall. A number of potential explanatory principles are considered, and their difficulties enumerated. It is concluded that the generation effect is real and that it poses an interesting interpretative problem.

This is an empirically oriented article whose purpose is to report a set of simple experiments that establish the existence of a robust and interesting phenomenon of memory. This phenomenon, called the generation effect, is robust in that it manifests itself across a variety of testing procedures, encoding rules, and other situational changes. It is interesting in that it does not seem to be easily or satisfactorily accommodated by any of the currently familiar explanatory notions. We expect that once the phenomenon is described in its initial form, it will be the subject of wider experimental analysis and will eventually become better understood.

In contrast to the usual objective reasons for embarking upon a line of research, the present work was neither initiated by any extant theoretical issue nor inspired by any previously published findings. It was carried out with the sole purpose of arriving at a clear answer to a straightforward factual question, namely, is a self-generated word better remembered than one that is externally presented? Most of us have probably encountered the informally expressed sentiment that there is an especial advantage to learning by doing, or that some kind of active or effortful involvement of the person in the learning process is more beneficial than merely passive reception of the same information. To what extent does this general notion have solid empirical support, as opposed to a casual or anecdotal base, particularly with respect to memory for self-generated verbal events versus those that have been read?

A search for some hard evidence in the journals uncovered no report of any thoroughgoing treatment of this question and no truly cumulative body of literature. There were a number of scattered references, some of which applied only tangentially to this problem, and although each was informative, they were characterized by such a diversity of methods, goals, and outcomes that no definitive overall conclusion could confidently be drawn from them as a whole (Abra, 1968; Anderson, Goldberg, & Hidde, 1971; Bobrow & Bower, 1969; Davies, Milne, & Glennie, 1973; Dosher & Russo, 1976;

This research was supported by National Research Council of Canada A7663. The experimental assistance rendered by Margaret Sparshott is gratefully appreciated.

Requests for reprints should be sent to Norman J. Slamecka, Department of Psychology, University of Toronto, Toronto, Ontario, Canada M5S 1A1.

Erdelyi, Buschke, & Finkelstein, 1977; Gardiner, Craik, & Bleasdale, 1973; Johnson, Taylor, & Raye, 1977; Russo & Wisher, 1976; Schwartz & Walsh, 1974; Underwood & Schulz, 1960; pp. 273–278). Further, some of the procedures were questionable, or else the results were, for one reason or another, not persuasive. For instance, conventional paired-associate learning has been compared with the case in which the subject freely provides his own responses (Abra, 1968; Underwood & Schulz, 1960, pp. 273–278). Such comparisons are hopelessly confounded not only by idiosyncratic item selection but by the fact that the conventional group has the advantage of continuous feedback on every trial, whereas the generation group necessarily does not. Again, Bobrow and Bower (1969) reported superior paired-associate recall with generated versus presented mediators, but a methodological refinement by Schwartz and Walsh (1974) saw the advantage completely disappear. Even the most relevant of these articles suffer from various disquieting complexities in their data. Thus, although Erdelyi et al. (1977) found no superiority of covertly generated words in the first two successive 5-min recall periods, they did observe a rise thereafter. Also, Anderson et al. (1971, Experiment 2) obtained unaccountably different outcomes between generate and read conditions, depending upon the particular test orders used.

In order to obtain an unbiased measure of the memorial consequences of generating versus being presented a word, it is critically important to avoid any possibility of confounding the effects of that variable with idiosyncratic item-selection habits which might confer an unfair advantage upon the generation condition. This requirement could most cleanly be met by employing the identical words for both conditions. To bring that about, the subject should not have free rein in generating but should be constrained in such a way that his responses are predictable beforehand and are the same as the words used for the presentation condition. The explicit form of constraint adopted for all the generation tasks in these experiments was as follows: The subject was given a rule, a stimulus word, and the initial letter of the response. He produced a word that began with the given letter and was related to the stimulus in the manner specified by the rule. To illustrate, with the rule *synonym,* the stimulus *rapid,* and the letter *f* the word *fast* would be generated. This task was readily handled by subjects and proved to be quite successful in reliably eliciting the desired responses, with errors of omission and commission being acceptably low. In the first experiment the effects of generation were assessed by recognition testing and confidence ratings, under five different rules.

Experiment 1

Method

Subjects and design. Twenty-four students of introductory psychology at the University of Toronto served on a voluntary basis and were given bonus credits toward their course grade for participating. Subjects for all succeeding experiments also came from this pool. The design was a 2 × 2 × 5 factorial. The main variable (*generate* versus *read*) was between subjects, as was the presentation rate (*timed* versus *self-paced*). The variable of *rules* was within subjects.

Materials. The input list consisted of 100 items, each on a separate index card. For the generate condition every card showed a stimulus word and the initial letter of the response, for example, *rapid–f.* For the read condition both words were present, for example, *rapid–fast.* There were 20 such items for each of five rules. The rules, with an example of each, were the following: associate (lamp–light), category (ruby–diamond), opposite (long–short), synonym (sea–ocean), and rhyme (save–cave). There were also several practice cards for each rule, which were used for instructional purposes at the start of the session to acquaint subjects with their task. The recognition test sheet was the same for all subjects and comprised 100 sets of three alternatives, each set having a response target and two new words as lures. The order of targets was random with respect to their original input sequence, and the lures of any set were not necessarily of the same apparent rule class as the target. For example, one set was *peer, diamond, critical,* with *diamond* the correct choice.

Procedure. Every subject was tested individually. After receiving task instructions and relevant practice cards, and just before the input phase, subjects were told that recognition of the responses would be tested at the end. Cards were

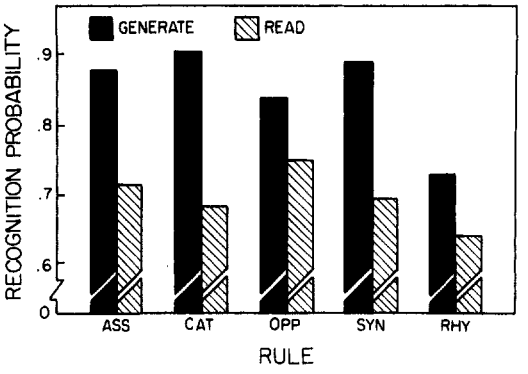

Figure 1. Mean recognition probabilities for each condition for each rule of Experiment 1. (ASS = associate; CAT = category; OPP = opposite; SYN = synonym; RHY = rhyme.)

blocked by rule, with order of rules varied across subjects. In all conditions a subject was told the operative rule and then handed the appropriate block of 20 cards. When he finished that block, he was told the next rule, handed the next block of 20, and so on. For every card the stimulus and response had to be uttered aloud just once, in that order. This equated the generate and read conditions on overt activity and also allowed the experimenter to monitor performance accuracy. In the timed condition, the subjects studied each card for 4 sec (the average rate for unpaced pilot subjects) and, when signaled by a timer tone, turned the card down and went on to the next. Self-paced subjects were to turn down each card as soon as a response was generated or read.

The recognition test was given immediately thereafter. Subjects used a cardboard mask that exposed only one set of alternatives at a time and proceeded in a fixed direction down the sheet without skipping or retracing. They were to encircle the one word in each set that occurred during the input phase and also to rate their confidence in each forced choice by using a 5-point scale from 1 (no confidence) to 5 (high confidence).

Results and Discussion

The overall median error rate in the generation task was only 6%, with 86% of the total originating with the associate and category rules. Items for which such errors occurred were eliminated from the scoring of the recognition data. Figure 1 shows the main recognition findings, with no correction for chance hits applied to these or any other recognition data in this article. For these and all subsequent statistical analyses, the alpha level was set at .05. Analysis confirmed that the substantial differences between generate and read conditions were highly significant, $F(1, 20) = 9.68$, $MS_e = .07$. On the other hand, the means for timed versus self-paced rates were .75 and .79, respectively, showing no difference ($F < 1$), nor did they interact with the generation variable. The main effect of rules was significant, $F(4, 80) = 5.28$, $MS_e = .01$. This was attributable solely to the lower overall recognition levels of rhyme responses. The interaction of generate versus read with rules was not significant, $F(4, 80) = 2.09$, $MS_e = .01$, indicating that the magnitude of the generation effect did not vary as a function of the particular rule involved. There were no other noteworthy findings.

Figure 2 displays the confidence ratings for correctly recognized items only. This way of reporting such data provides the more stringent and meaningful analysis, since the observed differences are not inflated by the presence of a correlation between confidence levels and sheer probability of recognition. Statistical tests revealed the very same pattern of outcomes that was obtained for recognition scores: Generate versus read was highly significant, $F(1, 20) = 6.92$, $MS_e = 1.04$; timed versus self-paced means of 3.95 and 4.05, respectively, did not differ ($F < 1$) and did not interact with generation; the effect of rules was significant, $F(4, 80) = 9.87$, $MS_e = .16$, and again there was no interaction of generate versus read with rules, $F(4, 80)$

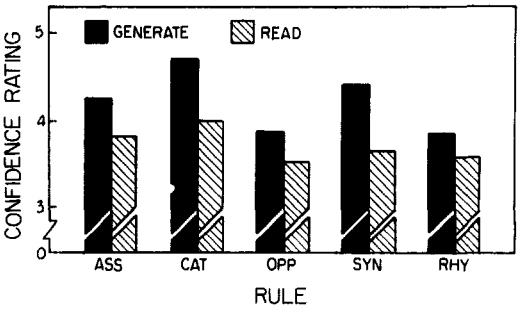

Figure 2. Mean confidence ratings for each condition for each rule of Expeirment 1. (ASS = associate; CAT = category; OPP = opposite; SYN = synonym; RHY = rhyme.)

$= 1.52$, $MS_e = .16$. There were no other significant effects.

These results are encouraging in that they vindicate the methodology adopted and suggest a positive answer to the question of whether there is a memorial benefit associated with the act of generating, as contrasted to just reading. Subjects who generated the words recognized more of them and were, in addition, more confident about doing so. Further, the effect persisted across a wide array of encoding rules as well as variation of input pacing. However, a single experiment does not establish a very wide or solid empirical base for any phenomenon. Additional questions arise about the stability of the effect under changed circumstances. For instance, would it still be obtained in a within-subjects design where each subject directly experienced the contrast between the two conditions? It has been demonstrated, for example, that the effect of pronunciation upon recognition performance depends critically upon whether a between- or within-subjects arrangement is employed (Hopkins & Edwards, 1972). A second question arises with respect to the fact that intentional learning instructions were given. Would the generation effect still emerge if subjects were not advised at the outset to prepare themselves for a subsequent recognition test of the responses? The next experiment addressed itself mainly to these two issues.

Experiment 2

Method

Subjects, design, materials, and procedure. Participants were 12 subjects from the same source as before. The design was a $2 \times 2 \times 5$ factorial with generate versus read as a within-subjects factor, informed versus uninformed about a test as a between-subjects factor, and rules, again, a within-subjects factor. The practice cards, all input cards, and the recognition test sheet were the same as those used for Experiment 1. Subjects were run individually. All had the same basic instructions and practice on the generate and read tasks. In addition, the informed group was told of the response recognition test to follow, whereas the uninformed group was not. The input cards were blocked by rules, with order of rules varied across subjects. Each rule block was divided into two subsets of 10 cards, one generate and one read. These subsets were so constituted that across subjects, all 100 items occurred equally often under both conditions. Furthermore, the order of conditions within blocks varied across blocks and across rules. Input was paced at a 4-sec rate. The subject was told the appropriate rule and then handed each subset of cards in succession. The generate and read tasks were executed in the very manner described for the prior experiment. The recognition test was the same forced-choice task as before, but instead of giving confidence ratings, the subject was asked to put either a *G* or an *R* next to each encircled item to indicate whether he had generated or read it. This permitted an estimate of the extent to which memorial discriminations could be made about the original input operations performed upon the correctly recognized items.

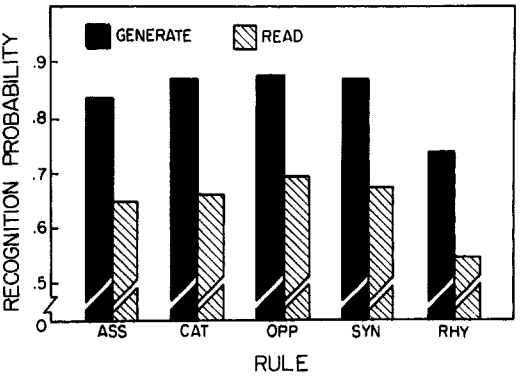

Figure 3. Mean recognition probabilities for each condition for each rule of Experiment 2. (ASS = associate; CAT = category; OPP = opposite; SYN = synonym; RHY = rhyme.)

Results and Discussion

The overall median error rate for the generation task was a modest 7%, with 77% of that total originating from associate and category rules; therefore, those items were eliminated from the recognition scores. Figure 3 shows the main recognition findings. The use of a within-subjects design brought out the generate versus read difference just as strongly as in Experiment 1, $F(1, 10) = 27.17$, $MS_e = .04$, a highly significant effect. The informed versus uninformed manipulation was totally inconsequential ($F < 1$), with means of .72 and .77, respectively. The effect of rules was once more significant, $F(4, 40) = 2.83$, $MS_e = .03$, again only because of the lower performance level for

rhyme words. There was no hint of any interaction between generate versus read and rules ($F < 1$). This latter finding reinforces the prior experiment's similar outcome in that regard by showing once more that the generation effect remained invariant across all of the encoding relationships examined. No other analyses were significant. Scores on the discrimination task for correctly recognized items revealed an overall accuracy rate of .74 for proper allocation of G and R responses. A test for the difference between that figure and a population mean of .50 was highly significant, $t(11) = 6.00$, $SE = .04$. Thus, recognition and correct allocation were related. As to the separate probabilities of correctly identifying recognized items as G versus R in origin, the data showed respective means of .70 and .77, which did not differ, $t(11) = 1.17$, $SE = .06$. For items falsely recognized, the allocation of G and R responses revealed a .44 − .56 split, suggesting no particular tendency toward bias or departure from equal use of the two categories.

The preceding data confirm the existence of the generation effect and further extend the range of circumstances under which it appears. Insofar as elicitation of the effect is concerned, it is immaterial whether a subject does or does not experience the contrasting conditions or does or does not have information about an impending recognition test. In addition, it is clear that correct recognition of an item is associated with a proper remembrance of the operations by which it was originally encoded. This, in general, is not particularly new or surprising, since there are many interesting examples of memory for ostensibly incidental aspects of remembered episodes (e.g., Geiselman & Bellezza, 1976), but it does help to strengthen the impression that generating and reading produce distinctively different memorial cues.

The next step in this research was to consider the potential effects of the generation task upon the stimulus members. Although the formal role of stimuli in this paradigm is only that of providing a source of constrain upon generation so that the responses are predictable, their presence nonetheless makes it possible to examine a particular question concerning the locus of the generation effect. It could be argued that the requirements of the generation task are such as to induce a heightened level of attention to all aspects of the situation and, specifically, that the stimulus member must also be attended more carefully than otherwise in order that it may effectively constrain selection of the response. In contrast, the reading task might not entail more than a superficial processing of the display, including the stimulus, since the response is provided in any case. If this surmise be correct, then it follows that the benefits of the supposedly more elaborate overall processing required for generation purposes would be reflected in superior memory on the stimulus side as well as on the response side. In that case the term *generation effect* would be a misnomer, since the stimuli are never generated in this paradigm. The third experiment investigated this possibility.

Experiment 3

Method

Subjects, design, and materials. Participants were 24 subjects, employed in a $2 \times 2 \times 2$ factorial design. The generate versus read variable was within subjects, the stimulus versus response recognition variable was between subjects, and a third variable of informed versus uninformed of a test was also between subjects. The input list consisted of 66 rhyme items, one per index card. There were also some practice cards to acquaint the subjects with the generation and reading tasks. The necessity for restricting the materials to rhymes arose because of the nature of the recognition test. It is not unlikely that stimulus recognition can be mediated by backward access from recallable responses. Nor is it unlikely that generated responses are more recallable than those that have been read. Therefore, the generate condition might do better on a stimulus test only as a consequence of its superior response accessibility and not by virtue of any intrinsic recognition superiority. This potential confounding was averted by always supplying the nontested member of the pair, thus equating its accessibility in both generate and read conditions. For example, a stimulus recognition item appeared as *wave, save, rave–cave,* and the corresponding response recognition item was *save–wave, cave, rave.* In order to have plausible lures for all 66 items,

rhymes were the obvious materials of choice; it would be next to impossible to provide equally good sets of lures for rules such as synonym or opposite. There were two recognition test sheets, one for stimuli and one for responses. Each consisted of 66 three-alternative sets and a cue, all rhyming. The ordering of targets was random with respect to their original input sequence.

Procedure. Subjects were individually tested. After basic instructions and practice on the two tasks, the informed group was told of a terminal recognition test on the pairs to follow, whereas the uninformed group was not. The input list was paced at a 4-sec rate, with generate and read items handled in the very manner as before. All subjects received the items in the same order. For half the subjects the first 33 items were generate and the last 33 were read, while the reverse was the case for the other half. Thus, each item occurred equally often in both conditions. Then, the appropriate half of each subject group was given one of the recognition sheets. Using a mask and moving it unidirectionally, they were to encircle the correct item and rate their confidence in each forced choice by referring to the same 5-point scale as used in Experiment 1.

Results and Discussion

With these rhyme materials, generation errors were almost nonexistent at .75%. The essential findings are displayed in Figure 4. Analysis of recognition probabilities showed a highly significant main effect of generate versus read, $F(1, 20) = 26.19$, $MS_e = 50.99$; a significant main effect of stimulus versus response, $F(1, 20) = 7.81$, $MS_e = 183.09$; and a highly significant interaction between them, $F(1, 20) = 10.75$, $MS_e = 50.99$, all indicating that the superior recognition of generated responses was not paralleled by a similar superiority on the part of the stimuli involved in the same task. The results of t test confirmed the existence of a large generation effect for responses, $t(20) = 5.94$, $SE = 2.92$, but none for stimuli. The informed versus uninformed manipulation was once again utterly inconsequential ($F < 1$), with means of .57 and .56, respectively. There were no other noteworthy results.

Analysis of confidence ratings for correctly recognized items showed a highly significant main effect of generate versus read, $F(1, 20) = 15.79$, $MS_e = .06$, and a highly significant interaction between the

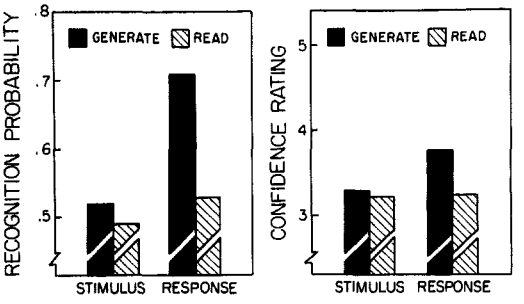

Figure 4. Mean recognition probabilities (left panel) and confidence ratings (right panel) for each condition and each type of test of Experiment 3.

latter and stimulus versus response, $F(1, 20) = 8.10$, $MS_e = .06$, again pointing to a generation effect that favors responses only. The results of t tests confirmed that impression by revealing a significant generation superiority on the response side, $t(20) = 4.82$, $SE = .10$, but not on the stimulus side ($t < 1$). No other outcomes were reliable.

The findings of this experiment are clear-cut. With respect to the question that initiated it, the answer is that for these materials, responses do show a generation effect, whereas stimuli do not. This conclusion holds both for recognition probabilities and for their associated confidence ratings. There is no support for the notion that the generation situation fostered heightened attention to all of the elements involved in it, with consequent memorial benefits accruing to them all. Rather, the effect was quite selective, falling only upon the generat*ed* element. Therefore, it seems that the designation originally applied to this phenomenon turns out not to be a misnomer—it is indeed a generation effect. This conclusion should be tempered by noting that it does not as yet cover semantically based relations. The latter may or may not act differently.

It was felt at this point that initial exploration of the phenomenon had now gone far enough through the use of recognition measures and that efforts should next be directed toward determining whether it also occurs with recall testing. It has been reported that some variables that influence

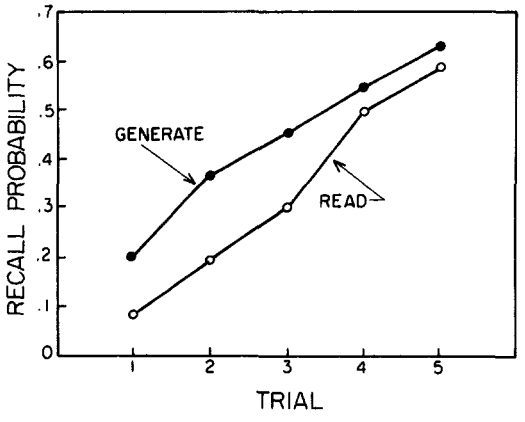

Figure 5. Mean free-recall probabilities for each condition for each trial of Experiment 4.

recognition performance do not necessarily influence recall in the same manner or to the same degree (see Brown, 1976; Tulving, 1976). It is therefore possible that the generation effect is limited to situations where copy cues are present and that the distinctive demands of a recall test might not bring it out. Accordingly, the next experiment focused upon multitrial free-recall learning of the response members.

Experiment 4

Method

Subjects, design, and materials. Participants were 12 subjects from the same source as before. A Subjects × Treatments design was employed, consisting of a 2 × 3 × 5 factorial with generate versus read, rules, and trials all as within-subjects factors. Since recall was to be tested, a shorter list would suffice. Therefore, only the synonym, opposite, and rhyme rules were used because they had proven to be relatively error free in generation of responses. There were 20 items for each rule, all on index cards, making a total input list of 60 events.

Procedure. Each subject was tested individually. Following the basic instructions and some practice in carrying out the two procedures, subjects were informed that free recall of responses would be tested thereafter. The input list was paced at a 4-sec rate with each task executed as in prior experiments. Rules were blocked, their order was counterbalanced across subjects, and they varied for each subject across trials. Each rule block was divided into two subsets of 10 cards, one generate and one read, handed to the subject in appropriate sequence. For every subject the generate–read order within a block was varied across blocks. These subsets were so constituted that across subjects, all 60 items occurred equally often under both conditions. Across trials a subject repeatedly had the same group of words to generate and a different group of words repeatedly to read. Five alternating presentation and test trials were administered. After each input trial there was a 30-sec period of backward number-counting to nullify short-term memory contributions. Every test trial called for a written free recall of responses on a blank sheet of paper, with 4 min allowed for this task.

Results and Discussion

The learning curves are displayed in Figure 5. As simple inspection suggests, there was a highly significant main effect of generate versus read, $F(1, 11) = 19.84$, $MS_e = .05$. Trials, as expected, was also very significant, indicating that learning took place, $F(4, 44) = 228.82$, $MS_e = .01$. In substantiation of the observed tendency for the two curves to draw together on the last two trials as learning progressed, the interaction of generate versus read with trials was reliable, $F(4, 44) = 4.47$, $MS_e = .01$. However, the interaction of the generate versus read with rules was not significant, $F(2, 22) = 1.22$, $MS_e = .10$. This latter result is consistent with the recognition data of Experiments 1 and 2 and reinforces the impression that the generation effect is constant or invariant across the kinds of encoding operations used here. The effect of rules was not reliable, $F(2, 22) = 1.41$, $MS_e = .05$, nor were there any other noteworthy findings.

Given that this was a free-recall situation, the possibility exists that the obtained generate versus read difference was somewhat

Table 1
Mean Output Serial Positions of Generated and Read Items for Each Trial, and Across All Trials, of Experiment 4

Condition	Trial					
	1	2	3	4	5	All
Generate	6.18	11.32	12.24	17.24	20.29	13.45
Read	5.72	8.36	13.76	16.30	19.56	12.74

exaggerated by the action of potential output interference. That is, whatever causes the superior accessibility of generated items may also cause those items to be recalled first, and so produce a decreased level of recall for the latter. If that occurred in this situation, generated words should be emitted in earlier output positions compared to those read. This is easily checked. Table 1 shows the mean serial position in output order for generate versus read items, trial by trial, and overall. Statistical analysis verified what inspection of the table already suggests, namely, that generate items were not favored with earlier recall either overall (no main effect) or within any trial (no interaction). Therefore, the superior performance of generated items as seen in Figure 5 may be interpreted as reflecting solely their great intrinsic accessibility.

The multitrial feature of this experiment also permitted a further type of analysis that was not possible with any of the preceding ones. The functional sources of the generation effect could be objectively allocated to two independent performance components, namely, the retention of old information from one trial to the next and the acquisition of new information between one trial and the next. An estimate of the contribution from the first source is the conditional probability of correctly recalling words on trial n that were also correct on $n-1$, and from the second source, the conditional probability of correctly recalling words on trial n given no recall on $n-1$. These two measures, respectively denoted as C/C and C/N (related to Tulving, 1964), were applied to the data, and the results are displayed in Figure 6. Inspection indicates that generation was superior to reading on both of these components across the first three trials but not thereafter, which is consistent with the tendency of the learning curves shown in Figure 5 to converge over the last two trials. Statistical analysis confirmed this general impression with a significant main effect of generate versus read, $F(1, 11) = 6.06$, $MS_e = .02$, and a highly significant interaction of that variable with trials, $F(3, 33) = 4.62$, $MS_e = .03$. The lack of any interaction of generate versus read with C/C versus C/N confirmed that the obtained generation superiority applied to both components. There were no other findings of interest.

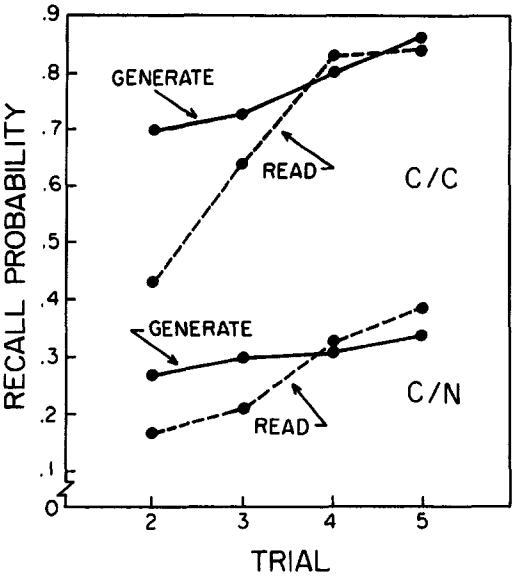

Figure 6. Mean C/C and C/N probabilities for each condition for each trial of Experiment 4, (C/C = conditional probability of correctly recalling words on trial n that were also correct on $n-1$; C/N = conditional probability of correctly recalling words on trial n given no recall on $n-1$.)

We can conclude that the generation effect reliably manifests itself not only with recognition but also under the unique demands of multitrial free-recall testing. This extends the effect's generality and shows that no externally provided retrieval cues are necessary in order to bring the phenomenon about. Further, it was seen that such recall superiority expressed itself both in better retention of items between trials as well as in better acquisition of new items within trials. The observation that generated words continued to enjoy an absolute advantage across successive trials in spite of the fact that for all trials after the first their generation was a repetitious act and no longer new is also to be noted. One interpretation of this could be that whatever is responsible for the generation effect does not habituate with repeated exposure to the

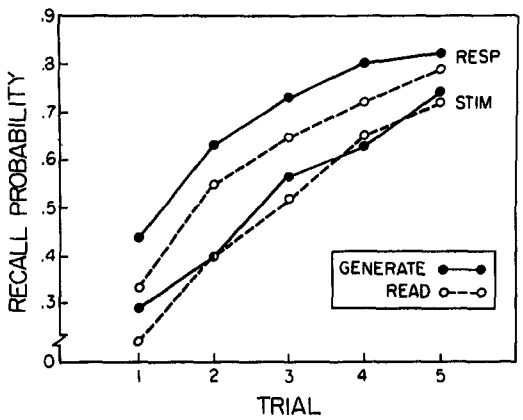

Figure 7. Mean cued-recall probabilities for each condition for each type of test of Experiment 5. (RESP = response; STIM = stimulus.)

same words but continues to augment their intertrial and intratrial retention. However, an alternative interpretation is that the advantage gained on the first trial was simply maintained thereafter by normal learning. But would that initial advantage have been lost if, on subsequent trials, subjects were switched from generating to reading those words? The present data are obviously not able to answer that interesting derivative question, but the problem should be reserved for future consideration.

The last experiment in this exploratory series addressed itself to the issue of whether the stimulus members involved in the generation task are also better able to be recalled. This is the counterpart of Experiment 3, in which the question was asked with respect to cued recognition testing. The fact that stimulus recognition was not enhanced in that experiment cannot simply be assumed to hold for recall as well, since the field's current theoretical grasp upon the nature of these two procedures is far from secure.

Experiment 5

Method

Subjects, design, materials, and procedure. Participants were 24 subjects from the same source as before. The design was a $2 \times 2 \times 5$ factorial with generate versus read as a within-subjects variable, stimulus versus response recall as a between-subjects variable, and the five trials as a within-subjects variable. The input materials (and practice cards) were the ones used in Experiment 3 and consisted of 66 rhyming pairs. There were two types of recall sheets, one for stimuli and one for responses, and every subject received five of one or the other type. Each item was cued with the nontested member, as *save*–____, or ____–*cave*. The order of items was random with respect to their original input sequence, and five different random orders were used across the five sheets. The decision to employ cued recall was predicated upon the same reasoning as in Experiment 3, namely, since generation facilitates response accessibility and since stimulus recall might be partially mediated through responses, the generate condition would enjoy a spurious advantage. Therefore, to equate the two conditions on functional accessibility of the untested member, it was always supplied.

All subjects were run individually. After receiving basic instructions and practice on the two tasks, they were informed that hereafter there would be a cued recall test of the stimuli (or responses). Input was paced at a 4-sec rate, and the tasks were carried out as in prior experiments. After each input trial there was a 30-sec period of backward number-counting, followed by a written cued-recall test for which 5 min were allowed. Five alternate training and test trials were given. Input items were divided into two blocks on every trial—one generate and one read—whose sequence was varied across trials and counterbalanced across subjects. The blocks were so constituted that across subjects, all items occurred equally often in both conditions, but for any given subject the same words were always to be generated or read on all trials.

Results and Discussion

The overall error rate for the generation task was a trivial .25%. All findings of the experiment are displayed in Figure 7. Analysis revealed a significant main effect of generate versus read, $F(1, 22) = 6.19$, $MS_e = .02$, a significant main effect of stimulus versus response, $F(1, 22) = 4.47$, $MS_e = .23$, and the expected large main effect of trials, $F(4, 88) = 120.57$, $MS_e = .01$, showing that learning had taken place. The interaction of interest, between generate versus read and stimulus versus response, was not reliable, $F(1, 22) = 1.82$, $MS_e = .02$. Although not supported statistically, Figure 7 shows that there was a larger absolute difference in favor of generate over read on the response side as compared to the stimulus side. Across trials, response recalls for

generate and read were .68 and .61, respectively, which represents a difference more than 3 times greater than that produced with corresponding stimulus recalls of .53 and .51. No other interactions in this analysis were reliable, either.

The above data replicate those of the preceding experiment in demonstrating a generation effect, which persisted across all trials of a multitrial recall learning task. Further, they extend the generality of the phenomenon to the cued-recall situation, with materials that call for only a single encoding rule throughout. The same list was previously tested with recognition in Experiment 3, and it is now quite evident that the effect does not at all depend upon the use of a *categorized,* or multirule, list of items. With respect to the specific question that gave impetus to this experiment, there was no significant interaction favoring recall of the response member as opposed to the stimulus member of generation pairs. This is clearly inconsistent with the presence of such an interaction in Experiment 3 and makes a straightforward answer to the question more difficult. Since the data were visually suggestive of an interaction, it might be prudent to defer any conclusion on the matter until a replication using more subjects is available.

General Discussion

These five experiments have clearly established the existence of the generation effect and have demonstrated it to hold across a fairly wide range of circumstances. The phenomenon was readily produced, at high levels of statistical reliability, even in experiments employing only 12 subjects. On the basis of data gathered up to this point, the following empirical conclusions may be stated. When a word was generated in the presence of a stimulus and an encoding rule, it was better remembered than when that same word was simply read under those conditions. The effect emerged with measures of free and cued recall, cued and uncued recognition, as well as confidence ratings. It applied to the whole gamut of encoding rules examined, and it persisted under multitrial learning requirements. Further, it was also invariant with respect to between- or within-subjects experimental arrangements, paced or unpaced presentations, and the presence or absence of instructions about a subsequent memory test. Finally, it was shown that the effect was limited to the generated word only and did not include the stimulus member of the display when testing was by cued recognition, but that a similar selectivity could not be claimed when testing was by cued recall. All the preceding findings constitute a reasonable start toward a delineation of the phenomenon and provide the necessary empirical base from which more analytic experiments can be launched in the future.

To the extent that the above data will meaningfully guide it, some theoretically oriented discussion is probably appropriate even at this early state of knowledge of the effect, in order to limn in a few of the explanatory possibilities as well as their attendant difficulties. Ideally, the phenomenon should be explained in the sense that it is seen to be simply another manifestation of some more general and overarching law of behavior. The next question is whether any well-founded principle is already available to do the job. Accordingly, several familiar notions, admittedly varying in their degrees of "well-foundedness," will now be considered.

That broad class of primarily quantitative formulations, which includes approaches such as strength theory, frequency theory, the law of exercise, and perhaps the total-time hypothesis—all of which share the predilection of attributing performance differences to the consequences of differing frequencies of situational occurrence (whether overt or covert) of events—can be rejected with some confidence. The present procedures always equated the generate and read conditions on the amount of overt responding to each card at input, and the pacing permitted no more covert rehearsal opportunities for the former group than for the latter. It could even be claimed that the read condition allowed more time for re-

hearsal because the response was fully accessible the moment the card was seen. The one place where this type of approach might make a contribution is in accounting for the maintenance of generation superiority across trials, as in Experiments 4 and 5. Since recalling an item increments its frequency and since the generation group started off with higher recall, its continued superiority may only be reflecting that original advantage. Albeit plausible, this is really tangential to the fundamental question of how that advantage came about in the first place. It is difficult to see how any of these basically quantitative representations can provide an acceptable answer, especially since the phenomenon has its basis in the qualitatively different activities of generating and reading.

Attention is next directed to a qualitative principle, namely, one that regards memorial differences as attributable to variations in the level or type of processing that items undergo during input (Craik & Lockhart, 1972; Hyde & Jenkins, 1969). The deeper or more elaborate the processing, the better the performance, all other things being equal. Deeper processing is semantic in nature, whereas shallower processing concerns itself with acoustic, visual, or other "superficial" features of the input. Attempts to apply this potentially useful notion to the present phenomenon encounter some problems of plausibility, as spelled out in the following three examples.

First, it can reasonably be maintained that the constraining stimulus is at least as much a recipient of the processing taking place during the generation task as is the response. Without an adequately analyzed stimulus, the appropriate response could not be generated. Indeed, the response can be viewed as an overt end product of the rule-guided stimulus-analyzing effort. Since the stimulus must necessarily be encoded to at least the same depth as the response which it elicits (how could a solely acoustically processed stimulus activate a semantically related response?), it follows that it should be as well remembered as the response. Why, then, did the observed memorial benefit fall only upon the response in Experiment 3 and not at all upon that other member which was also processed to a comparable extent? Although that outcome is not easily reconciled with the principle under consideration, the question of stimulus memorability is still open because of the findings of Experiment 5. Second, this view commits itself to one relevant testable prediction concerning a rule-related effect upon memory. The rhyme rule should produce only a relatively shallow level of processing as compared to the other rules, with a consequent attenuation or even absence of a generation effect in its case. Again, the data do not bear this out. In none of these experiments was the rhyme rule ever singled out as having mediated a significantly weaker effect, nor was an interaction observed between rules and generate versus read. The stability of the effect hardly invites a "levels" explanation. Another prediction, although irrelevant to the central question, calls for a main effect of rules, with rhyme expected to show the lowest overall memory performance. This was confirmed by the recognition data of Experiments 1 and 2, but it still does not account for the generation effect as such, which occurred within all rules. Third, and most speculative, it might be surmised that the act of generation itself, regardless of what encoding rule applies, intrinsically entails a more profound processing level than does the virtually automatic act of reading. This suggestion represents a novel and untested extension of the usual domain of situations embraced by the levels-of-processing literature. In the absence of any independently determined prior assessment of the processing depths characteristic of acts of generation versus reading, such an explanation is clearly post hoc and would first require some validating experiments to give it substance. It remains to be determined whether an approach predicated upon qualitative differences in encoding can successfully deal with these problematic aspects of application.

A third notion, which is really a subset of the more general principle considered above,

focuses upon the paired-associates structure of the present materials. It would say that the generation task forces a distinctive encoding of the relation between stimulus and response, in contrast to the reading task, which does not effectively demand any registration of that relation. In spite of the consistent practice of informing both conditions of the operative rule, it is still possible that the read items did not encourage use of that information (since it wasn't necessary), with the result that their encodings lacked relational specificity. To the extent that such distinctiveness is a factor in memory, the generation effect might be accommodated. This notion is evidently a salient one, since it occurred independently to the authors and to an editorial reader. Experiments that test it are already under way and will be reported later.

A fourth possible avenue of interpretation will be described, which appeals to a principle that probably has the least established status but is of considerable interest nonetheless. It is the idea that an initial recall task confers beneficial consequences upon a subsequent memory test on the same material. As applied to the present paradigm, it would stress that the act of generation is really an instance of recall, with the source being semantic memory. The subject neither learns nor creates anything new, but simply retrieves an existing item of information from his repertoire of knowledge, guided by the cues of stimulus, encoding rule, and initial letter. The resulting overt response constitutes the episode that is later tested for retention. In contrast, the reading task involves no recall-based episodes, since all responses are given. Superior memory for generated items is then referred to the fact that they enjoyed a prior recall (generation), which somehow served to increase their subsequent memorability. At first glance this also appears to be a promising account, provided that its basic premise has adequate support from the literature. It seems clear enough that prior recall activity improves subsequent recall when compared either to reading the material (Gates, 1917) or to engaging in a filler task (Darley & Murdock, 1971). However, evidence about the influence of prior recall upon subsequent recognition is not unequivocal, with absolutely no benefit found by Darley and Murdock (1971) but positive effects reported in some other cases (Broadbent & Broadbent, 1975; Lockhart, 1975). In addition, there is one feature of those studies that deviates sharply from the current procedure, namely, the fact that all items were presented by the experimenter, whereas the current paradigm contrasted externally originated events with entirely self-produced ones. Can the memorability of an episode created by recalling from semantic memory be legitimately equated with the dual episodes involved after recalling an event of external origin? Perhaps not (for example, Slamecka, 1966). This type of explanation also comes close to being only a restatement of the very findings it seeks to explain, that is, that a generated word is better remembered than one that was read because it was generated (recalled). One substantive test of this notion might be to determine whether *failure* to generate a word (with the correct response then provided) still results in better memory than in the simple read condition. If so, it could not be attributed to a prior recall. This is reminiscent of the Gardiner et al. (1973) experiment, which, however, lacked a read comparison group.

These various interpretations of the generation effect are not exhaustive but only indicative of the variety of approaches possible. Others could also be advanced: Generation requires more cognitive effort than does reading, and effort increases memorability; generation entails extensive tagging of nodes in the associative network, thus increasing access routes; generation is response emission without copy prompts, which is the best training for a later test. All of these offerings are speculative at this time in the sense that the available data do not suggest a rational preference among them. The present experiments were designed with the modest intent of delineating the outlines of this phenomenon and not of testing theoretical hypotheses concerning its

cause. It would be premature, therefore, to press for any one of these representations until the appropriate analytic studies are carried out.

References

Abra, J. C. Acquistion and retention of consistent associative responses with varied meaningfulness and similarity of stimuli. *Journal of Verbal Learning and Verbal Behavior,* 1968, *7,* 647–652.

Anderson, R. C., Goldberg, S. R., & Hidde, J. L. Meaingful processing of sentences. *Journal of Educational Psychology,* 1971, *62,* 395–399.

Bobrow, S. A., & Bower, G. H. Comprehension and recall of sentences. *Journal of Experimental Psychology,* 1969, *80,* 455–461.

Broadbent, D. E., & Broadbent, M. H. P. The recognition of words which cannot be recalled. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V.* New York: Academic Press, 1975.

Brown, J. An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition.* London: Wiley, 1976.

Craik, F. I. M., & Lockhart, R. S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior,* 1972, *11,* 671–684.

Darley, C. F., & Murdock, B. B., Jr. Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology,* 1971, *91,* 66–73.

Davies, G. M., Milne, J. E., & Glennie, B. J. On the significance of "double encoding" for the superior recall of pictures to names. *Quarterly Journal of Experimental Psychology,* 1973, *25,* 413–423.

Dosher, B. A., & Russo, J. E. Memory for internally generated stimuli. *Journal of Experimental Psychology: Human Learning and Memory,* 1976, *2,* 633–640.

Erdelyi, M., Buschke, H., & Finkelstein, S. Hypermnesia for Socratic stimuli: The growth of recall for an internally generated memory list abstracted from a series of riddles. *Memory & Cognition,* 1977, *5,* 283–286.

Gardiner, J. M., Craik, F. I. M., Bleasdale, F. A. Retrieval difficulty and subsequent recall. *Memory & Cognition,* 1973, *1,* 213–216.

Gates, A. I. Recitation as a factor in memorizing. *Archives of Psychology,* 1917, *6,* No. 40.

Geiselman, R. E., & Bellezza, F. S. Long-term memory for speaker's voice and source location. *Memory & Cognition,* 1976, *4,* 483–489.

Hopkins, R. H., & Edwards, R. E. Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior,* 1972, *11,* 534–537.

Hyde, T. S., & Jenkins, J. J. Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology,* 1969, *82,* 472–481.

Johnson, M. K., Taylor, T. H., & Raye, C. L. Fact and fantasy: The effects of internally generated events on the apparent frequency of externally generated events. *Memory & Cognition,* 1977, *5,* 116–122.

Lockhart, R. S. The facilitation of recognition by recall. *Journal of Verbal Learning and Verbal Behavior,* 1975, *14,* 253–258.

Russo, J. E., & Wisher, R. A. Reprocessing as a recognition cue. *Memory & Cognition,* 1976, *4,* 683–689.

Schwartz, M., & Walsh, M. F. Identical subject-generated and experimenter-supplied mediators in paired-associate learning. *Journal of Experimental Psychology,* 1974, *103,* 878–884.

Slamecka, N. J. Differentiation versus unlearning of verbal associations. *Journal of Experimental Psychology,* 1966, *71,* 822–828.

Tulving, E. Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychological Review,* 1964, *71,* 219–237.

Tulving, E. Ecphoric processes in recall and recognition. In J. Brown (Ed.), *Recall and recognition.* London: Wiley, 1976.

Underwood, B. J., & Schulz, R. W. *Meaningfulness and verbal learning.* Philadelphia: Lippincott, 1960.

Received February 20, 1978
Revision received June 9, 1978 ■