Journal of Experimental Psychology
1974, Vol. 103, No. 4, 751–757

DEPTH OF PROCESSING PICTURES OF FACES AND RECOGNITION MEMORY [1]

GORDON H. BOWER [2] AND MARTIN B. KARLIN

Stanford University

These studies ask whether *S* remembers a picture better the greater the "depth of processing" he allots to it. Depth of processing pictures of faces was varied according to judgments of sex ("superficial") or judgments of likableness or honesty of the person pictured. Performance on a later recognition memory test was high for pictures judged for likableness or honesty, and low for pictures judged for sex. This ordering held as true for intentional learners as for incidental learners. A final experiment showed that face recognition memory was not materially affected by a context manipulation: an *old* test picture was remembered at a level determined by its original depth of processing and independently of how it was tested—either alone, along side an *old* picture it had been studied with, or with a *new* picture.

In a recent paper, Craik and Lockhart (1972) argued that "depth of processing" of stimulus material is a direct determinant of how well that material will be remembered. The underlying assumption in their approach is that a stimulus is processed through a series of stages with different kinds of information being extracted from or triggered off by the stimulus at successive stages. Sensory features of the stimulus are presumably extracted first, whereas associative information (such as the name or meaning of a grapheme) becomes available later. In support of their depth of processing hypothesis, Craik and Lockhart review studies showing higher incidental learning for words which *S*s had processed for meaning than for items processed for physical attributes. For instance, Hyde and Jenkins (1969) oriented *S*s to answer different questions with respect to a word, either counting the number of letters in it or the number of *e*s, or rating it for pleasantness. Those *S*s who did the pleasantness judgments recalled the words later much better than did the other *S*s. Similarly, Johnston and Jenkins (1971) showed that *S*s required to think of an adjective appropriate to a presented noun had greater recall for the nouns later than did *S*s required to supply a rhyming word for the noun. To gain further support for the hypothesis, Craik (1973) had *S*s make one of three judgments about a presented word: whether it (*a*) was printed in upper- or lowercase letters, (*b*) rhymed with a given word, or (*c*) made sense when inserted into a given sentence frame. Craik found that the reaction time for these judgments increased in the order given above, and that recognition and recall on a later unexpected memory test increased linearly with reaction time (RT). Presumably, it is not the decision RT per se that is important, but the extent of semantic processing of the item.

It is noteworthy that all the studies supporting the depth of processing hypothesis have used words as learning materials. There is therefore a need to explore the generality of such effects with other stimulus domains. Memory for pictures naturally suggests itself as a possible counterexample, since such memory seems quite substantial even with relatively rapid presentation rates and low processing levels (see Haber, 1970; Shepard, 1967). Moreover, if some kind of "imagery code" for the item is set up within a few seconds of its presentation, then there may be nothing "deeper" to be established as a result of an orienting task.

Now, recognition memory for common pictures is notoriously high, so one must

[1] This research was supported by Grant MH-13950-07 to the first author from the National Institute of Mental Health.

[2] Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, California, 94305.

find some class of pictures which will lead to less spectacular memory. For this reason, we used pictures of faces drawn from a college student yearbook as learning materials; these faces are relatively homogeneous and *S*s show substantial forgetting of them over intervals of several minutes. The orienting tasks studied here involved judgments by *S*s with respect to three dimensions of the picture: sex, likableness, and honesty of the person in the picture. Judgment of sex presumably represents a lower level of processing than the other two judgments, so according to the depth of processing hypothesis, it should result in lower recognition memory than the other two. There was no a priori reason for ordering the other two judgments.

EXPERIMENT I

Method

Subjects and design. The *S*s were 12 university students who were fulfilling a participation requirement for their introductory psychology course. The design was within-*S*s, with repeated recognition measures obtained from each *S* at each of the three different levels of processing.

Materials. Two hundred-sixteen black-and-white 35-mm. slides were made from pictures in the 1972 *Yale Beacon*, the Yale University senior yearbook. These comprised 72 duplicate pairs along with 72 unique slides. The study set was made up of one member of each pair of duplicates. The test set was made up of the other member of each pair mixed in among the 72 unique slides, the "distractor" items. Half of the slides were of females, half were of males. Pictures were so selected as to insure some uniformity of the pictures, so distinctive cues were not readily apparent. Any picture of a person with any unusual characteristics (e.g., a hat, large earrings, etc.) was eliminated. All pictures were of Caucasians, and all of the males were wearing ties. The 72 study items were randomly ordered, divided into 6 groups of 12 slides each, and then placed in a slide tray with a blank slide separating each group. The 144 test items (the 72 duplicates plus 72 distractors) were randomly ordered and placed into two more slide trays with 72 slides in each. All slides were shown on a Kodak Carousel projector which advanced at the rate of one slide every 5 sec.

Procedure. The *S*s were led to believe that we were interested in how fast they could make different kinds of judgments about a person, using only his photograph as evidence. They were to indicate their judgment by pressing one of two buttons on a console before them. Wires from the console led through a hole in the wall into an adjoining room. Before starting the experiment, *E* excused himself to "turn on some data-recording equipment in the other room." Actually, the console buttons were not connected to any equipment. The cover task was furthered by (*a*) asking *S* if he had been "in any reaction time experiments before," and (*b*) instructing him to "be sure and press the button firmly so that the responses are recorded properly."

The *S*s were then informed of the three types of judgments. Each slide received one of three binary judgments, based on either sex, likableness, or honesty of the person shown. For the sex judgment, one button was to be pressed for *males*, the other for *females*. For the likableness judgment, *S*s were told to "use whatever criteria you want" to judge the degree of likableness of the person in the photograph, and to use one button for those people who were *more likable* than average, and the other for those who were *less likable*. Honesty was to be judged on a similar subjective scale, with one button for *more honest* and the other for *less honest*. The *S*s were told that as each slide was presented, they were to make their judgment as quickly and as accurately as possible. After their response, they were instructed to continue looking at the slide until the next one appeared. Each slide was presented for 5 sec. The *S*s in Experiment I were *not* told that they would have to remember the faces for a subsequent task.

Before each block of 12 slides, *S*s were told which judgment to make during that block. With three different judgments and six blocks of slides, each type of judgment was made twice, once during the first three blocks and once during the last three. The order of judgment types during the study set (as well as the assigned buttons) was counterbalanced across *S*s.

After the study set was shown, *S*s were informed that they would then view a set of test slides, some of which were duplicates of those they had just seen. For each slide presented, *S*s were to indicate on a response sheet whether they thought the slide had been in the study set (was *old*) or had not been studied (was *new*). The *S*s responded by circling a number from 1 through 6 on a response sheet. The number circled indicated the binary *old–new* decision along with *S*'s degree of confidence in the correctness of his judgment (from *guess*, to *probably*, to *sure*). The 144 test slides were then shown at a 5-sec. rate. After the recognition task was completed, *S* was debriefed.

Results

The primary data for the three conditions are shown in the left columns of Table 1, with Column 1 reporting the mean confidence ratings (where 1 = *sure old*) and Column 2 reporting the proportion of *old* responses to test stimuli. Responses to the distractors appear in the

TABLE 1

CONFIDENCE RATINGS (CR) AND HIT RATES (HR) FOR INCIDENTAL AND INTENTIONAL LEARNERS

Judgment	Experiment I—incidental learners		Experiment II—intentional learners	
	CR	HR	CR	HR
Honesty	2.19	.81	2.41	.76
Likableness	2.37	.75	2.08	.80
Sex	2.98	.60	3.09	.56
Distractors	4.73	.15[a]	4.98	.12[a]

[a] False positive rate for distractors.

bottom row; there was a 15% "false positive" error rate to the new distractors in Experiment I. The standard deviations for the confidence ratings are all near .42, and for the *old* proportions are all near .10.

Clearly, *S*s show good discrimination between any *old* stimulus and a *new* stimulus. Also, pictures that had been judged for sex were remembered less well than pictures judged for honesty or likableness. This contrast was significant at the .001 level for confidence ratings and for hit rates (arc sine transformation). A further contrast compared memory for pictures receiving honesty vs. likableness judgments. A correlated t test declared the advantage of the honesty judgment to be marginally significant for both confidence ratings, t (11) = 2.38, $p < .05$, and hit rates, t (11) = 3.05, $p < .02$.

Discussion

A few procedural remarks are appropriate. First, the possible "ceiling effect" for picture recognition was avoided by use of face pictures, which are a readily available, homogeneous source of stimuli. Over a retention interval of 15 min. or so, the recognition rate did not exceed 81%. Second, the task deception was effective: no *S* expected the recognition memory test, and all found the judgment tasks reasonable and interesting.

We have presumed that our three orienting tasks require differing degrees of processing or analysis of the pictured face. Evidently, we may conclude that these differing degrees of processing of a picture produce corresponding differences in later memory for that face. What is missing in this description is a fuller characterization of what depth of processing means with faces.

We may compare our results with those of Craik (1973) or Hyde and Jenkins (1969), who showed better recall or recognition of words when the orienting task required *S*s to process the semantic meaning of the word rather than its visual appearance. Although a face hasn't a "semantic meaning" in the sense that a word does, it still may trigger a number of associations, such as the resemblance of the face to that of a friend or celebrity. Thus, greater depth of processing of a face might correspond to a greater number of unique associations that *S* retrieves from his memory. Alternatively, we might describe the depth of processing of faces in terms of the amount of detail extracted during analyses when comparing the stimulus to an array of prototypes or set of criteria for deciding upon such vague classifications as honesty or likableness. We may suppose that a fraction of these extracted features are represented in the memory trace of the face; the later recognition memory just reflects this difference in number of descriptive features stored.

A plausible alternative to the above hypotheses is that because sex judgments can be made more quickly than honesty or likableness judgments, *S* spends less time per se looking at the faces during study. Therefore, the later difference in recognition memory would, on this hypothesis, simply reflect the difference in *functional* (effective) exposure time to the picture to be learned. To assess this alternative account, Experiment II used the same three processing tasks, but *S*s received intentional learning instructions. In this case, it would be expected that *S*s would study the face for the full time it is in view.

EXPERIMENT II

Method

The procedure was identical in all respects to the preceding study, except for the addition of intentional learning instructions along with the orienting judgment (sex, honesty, or likableness). Twelve new *S*s from the same source were told to "study each picture carefully even after making your response so that you will be able to recognize the pictures in a subsequent task."

Results

The data were analyzed as before and are summarized in the right-hand columns

in Table 1. As before, memory is poorest for pictures judged for sex; comparing it to the honesty and likableness conditions combined, t (11) = 7.52, $p < .001$, for confidence ratings, and t (11) = 6.32, $p < .001$, for hit rates. As can be seen from Table 1, honesty and likableness judgments in this experiment were in reversed order from Experiment I. These two conditions showed a significant difference in confidence ratings, t (11) = 2.45, $p < .05$, but not in hit rate, t (11) = .78.

The recognition rates for intentional learners (right columns of Table 1) and incidental learners (left columns) are quite similar. Cross-experimental comparisons of recognition rates for the four classes of test stimuli yielded no significant differences. This finding of no difference appears to be a common one with recognition memory: the mode of processing the input is critically important, whereas the intention to learn or the expectation of a retention test appears of lesser importance.

The conclusion is that the effect of the depth of processing upon memory for faces is not a simple artifact of variation in functional study time. Rather, different amounts or different types of information are being stored about the face, depending on its mode of processing.

Experiment III

The final experiment searched for an influence of context upon face recognition. Our experiment is modeled after ones by Tulving and Thomson (1971) and Thomson (1972); they had *S*s study words presented singly or in pairs, and then tested their recognition memory for a given word presented either singly, in an *old* pair, or in a *new* pair. They found that any change in the surrounding context of presentation for a word reduced its recognition rate. They interpreted their results in terms of the encoding specificity principle: that the context prevailing at presentation influences how a stimulus is encoded or represented internally, and the retrieval conditions during testing must reactivate that old encoding in order for recognition memory to occur.

We asked whether we could show effects on face recognition by such elementary manipulations of presentation conditions. The Tulving and Thomson (1971) paradigm was combined with an intentional learning situation at two levels of processing. The orienting tasks involved judgments of either sex or compatibility. Judgments were made for pairs of faces presented simultaneously. For the sex judgments, each member of the pair was judged independently. In the compatibility condition, however, each picture was to be considered in relation to the other member of the pair. The *S*s had to judge whether the pair of people would be socially compatible with one another. This explicit comparison and relative evaluation process at input would seem to maximize the chances of finding a context effect on recognition memory. On these grounds, then, one would expect lower recognition for items under changed test contexts if they were originally judged for compatibility, but not if they were judged independently for sex. As before, we expected too that compatibility items will be recognized better than sex items.

Method

Subjects and design. Twelve *S*s (7 male, 5 female) participated in the experiment. Of these, 2 received course credit, whereas 10 were paid $1.75 for an hour's service. The experiment had a within-*S*s design, with repeated measures from each *S* for the 11 recognition conditions outlined below.

Materials. The slides were identical to those used in the previous studies, but they were presented in pairs. Two projectors showed slides simultaneously side by side, advancing every 8 sec. The study set consisted of the same 72 slides as used before, except they were presented as 36 pairs of slides by projecting each pair side-by-side on the wall of the experiment room. These 36 pairs were separated into two blocks of 18 pairs each. One block was judged for sex, the other for compatibility.

The test set again consisted of the duplicates of each study slide, plus the 72 distractors, presented in a new manner. In the recognition test, *S* saw the slides either singly or in pairs. When only one slide was to be viewed at a time, a dummy slide was placed in one slide tray so that half of the screen remained dark.

There were several different types of test items. *New* ("distractor") slides were presented singly ($n = 12$), with another *new* slide ($n = 40$), or with an *old* slide ($n = 20$). *Old* slides were presented alone ($n = 12$), beside the same *old* picture they had been paired with during study ($n = 20$), beside a different but *old* picture ($n = 20$), or beside a *new* slide ($n = 20$). Half of each type of *old* slide had been judged for sex and half for compatibility.

Procedure. The *S*s were told that they would have to make one of two decisions about the pairs of slides shown. For the sex condition, they were to judge the sex of each member of the pair and make *two* responses for each pair shown, the first response for the picture on the left, the second for the picture on the right. As previously, *S* responded by pressing dummy buttons on a console—one button for *male*, the other for *female*. For the compatibility condition, *S* was told to examine both pictures and judge how compatible the two people were, i.e., "whether or not they would be friends." Only one *yes-no* response was made for such a pair.

Each *S* made either the sex or compatibility judgment on all slides in the first block of 18 pairs, and the other judgment on the second block of 18. The order of conditions was counterbalanced across *S*s. The *S*s saw each pair of slides for 8 sec.; they were told to study the slides after making their responses, so that they would be able to recognize the pictures in a subsequent memory test.

After presentation of the study set, *S*s were given an answer sheet and the recognition task was explained. The same 6-point rating scale as before was used. They were instructed to make recognition judgments of each picture independently of the other one being shown. The *S*s responded by writing a number for each slide in the appropriate space on the answer sheet. The *S*s were told that *old* pictures would always appear on the same side of the screen that they had appeared on during the study set. During the recognition test, the slide pairs were shown every eight sec.

Results

Mean confidence ratings and hit rates were calculated for all 11 test conditions. Since these two measures yielded parallel results, only those for confidence ratings will be reported. These means are shown in Figure 1, with the various test conditions marked along the horizontal axis. Figure 1 conveys the main facts yielded by the statistical analysis. There is significant discrimination of *old* from *new* stimuli, and pictures involved in pairs judged for compatibility during study are

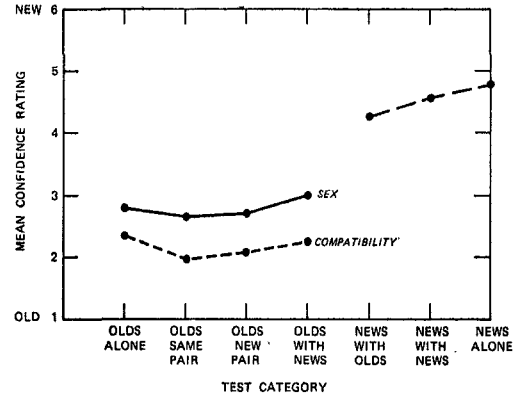

FIGURE 1. Mean confidence rating as a function of test category. (Note that for "Olds," the lower the rating, the better the performance.)

remembered much more than pictures judged for sex. For this latter contrast, t (11) = 2.59, $p < .05$, replicating our earlier results for face memory depending on depth of initial processing.

It is also apparent that there is little if any "context" effect upon these recognition judgments: The mean confidence rating assigned to an *old* picture does not vary much with the manner of its test presentation. The critical comparison is between performance to *old – same pair* (see Figure 1) which maintains the study context, and the three changed contexts for testing an *old* item. Combining the three latter conditions, their contrast to the *old – same pair* performance yielded the following four t values (all $df = 11$): sex stimuli—confidence ratings, $t = 1.16$, transformed hit rate, $t = 1.30$; compatibility stimuli—confidence ratings, $t = 2.07$ ($p < .10$), transformed hit rate, $t = 1.50$. Thus, there is no statistical evidence for a strong effect of test context upon recognition memory for *old* pictures. The largest context effect in the experiment occurs for performance to *new* stimuli (right side of Figure 1), where a new stimulus tested with an *old* one received a higher "recognition" rate (and lower confidence rating) than a new stimulus presented alone. This is the sort of result to be expected either if there is some "generalization of familiarity" among members of a test pair or a persisting "response set"

to say *yes* to both test cues when one item is definitely recognized. However, these trends in the *new* stimulus ratings fall short of statistical significance, so they should not be given much weight.

General Discussion

Two issues are to be discussed. The first concerns the lack of context effect upon face recognition memory in Experiment III. This occurred despite the use of the Tulving and Thomson (1971) paradigm, which did yield strong context effects with words as stimuli. The obvious interpretation is that in this task, emphasizing individual recognition of unambiguous faces, the encoding of each face was unaffected by its context. This might be expected on the basis that encoding faces is a highly routinized skill, and its output is not subject to much variability when the picture is unambiguous. Of course, one can compose ambiguous face pictures (e.g., the *wife vs. mother-in-law* picture) for which contextual settings would begin to operate for encoding and recognition memory.

There is a second context effect in face recognition familiar to all of us, that in which we fail to recognize an acquaintance when we come across him in some unexpected place. However, that effect would appear to be a case of the well-known effect of expectation upon perceptual achievements. That is, perceptual identification depends jointly on the strength of the sensory evidence and the expectations suggested by the background context (e.g., Morton's 1969 "logogen" model). In such an analysis, the familiar phenomenon of "I couldn't recognize you out of context" would have a different explanation than the Tulving and Thomson results showing encoding variability of single words presented for recognition.

A second issue for discussion concerns speculations about what depth of processing means for faces, and why it helps memory. Consider say, honesty judgments vs. sex judgments of a face. Now, for the present set of pictures, sex could be judged by noting one or another salient cue, such as presence of a necktie, short hair, rough skin, bushy eyebrows, cosmetic makeup, and so on. Judgment of honesty of a face would appear to require comparison to an idiosyncratic set of vague prototypes or criteria, regarding the patterning of features such as distance between eyes, size of eyes, size of pupils, curvature of mouth, thickness of lips, and so on. The decision is probably influenced by *S* remembering a person he knows who resembles the pictured person, with *S* then judging the honesty of the "stimulus person" according to his judgment of the "person retrieved from memory." There are probably further strategies by which such honesty judgments might be made (e.g., sampling "honest people I've known" from memory for comparison to the test stimulus). But suffice it to say that all such judgments seem to require that more features of the face be examined, that they be examined more times per second, and that there be more use of the stimulus to search through memory and compare to retrieved "honest prototypes." This differential processing of the face might be indexed by the number of eye fixations, since Loftus (1972) found greater recognition memory for pictures (of naturalistic scenes) that received more eye fixations during a constant study interval.

The constructionist view of memory (Neisser, 1967) would suppose that part of what is stored about a stimulus is the series of cognitive operations or constructive activities needed to arrive at a decision about the stimulus. An implication for the present case would be that if *S* "recognized" a test face, then he should be able to remember what kind of judgment (sex or honesty) he had made about it during the study trial. However, such data were not collected in the present experiment, since the "blocked" nature of the alternating judgment tasks would make the memory for judgment type correlate with memory for the position of an item in the study list. A study to unconfound list position and mode of processing is yet to be done.

An alternative prediction is that if *S* is asked during the recognition task to make either the same or a different judgment about each picture, performance will be facilitated when *S* makes the same judgment as during study, since he will repeat the same cognitive operations, i.e., access the same "schemata" as before.

To reiterate, the simple message of these experiments is that storage in memory of perceptions of faces (surely an overpracticed skill) can still be varied by requiring differing "depths" or degrees of analysis and processing of the face with respect to different criteria. The practical prescription is that if you want to remember a person's face, try to make a

number of difficult personal judgments about his face when you are first meeting him.

REFERENCES

Craik, F. I. M. A "levels of analysis" view of memory. In P. Pliner, L. Krames, & T. Alloway (Eds.), *Communication and affect: Language and thought.* New York: Academic Press, 1973.

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.

Haber, R. N. How we remember what we see. *Scientific American*, 1970, **222** (5), 104–112.

Hyde, T. S., & Jenkins, J. J. The 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.

Johnston, C. D., & Jenkins, J. J. Two more incidental tasks that differentially affect associative clustering in recall. *Journal of Experimental Psychology*, 1971, **89**, 92–95.

Loftus, G. R. Eye fixations and recognition memory for pictures. *Cognitive Psychology*, 1972, **3**, 525–551.

Morton, J. The interaction of information in word recognition. *Psychological Review*, 1969, **76**, 165–178.

Neisser, U. *Cognitive psychology.* New York: Appleton-Century-Crofts, 1967.

Shepard, R. N. Recognition memory for words, sentences, and picture. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 156–163.

Thomson, D. M. Context effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 497–511.

Tulving, E., & Thomson, D. M. Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology*, 1971, **87**, 116–124.

(Received February 21, 1974)