Bulletin of the Psychonomic Society
1979, Vol. 13 (1), 15-18

Does fluency of face description imply superior face recognition?

ALVIN G. GOLDSTEIN, KAREN S. JOHNSON, and JUNE CHANCE
University of Missouri, Columbia, Missouri 65201

While viewing a set of 10 standard portraits, independent judges used a checklist, composed of facial features, each with several appropriate descriptive adjectives, to construct verbal descriptions of the faces. Subjects also briefly viewed the standard portraits and used the checklist to construct verbal portraits from memory. Subjects' accuracy of description was assessed by comparing their responses to composite descriptions derived from the judge's concordant responses. In the final phase of the study, subjects' recognition memory for faces was measured on a completely new set of faces. Correlational analysis revealed no association between describing ability and recognition memory performance. Further analysis indicated that better describers might be slightly better recognizers, but the effect was weak. Two earlier studies also failed to demonstrate a relationship between recognition ability and describing ability.

When a witness to a crime is asked by police to describe what the culprit looked like, it is reasonable to suppose that, on the basis of the witnesses' interview behavior, the police form judgments about the accuracy of the description. A complete verbal portrait, including details of size, shape, and color, spontaneously elicited from a self-assured witness who appears to possess a clear mental "picture" of the culprit's face would be expected to impress police quite differently than would a sketchy verbal report with few descriptive details. Policemen probably assume that the fluent eyewitness, in contrast to the taciturn witness, is more likely to recognize the culprit from either a photograph or a face-to-face encounter. The present research addresses itself to the question, in what way is the ability to accurately describe faces in verbal terms related to face recognition performance?

This problem was also explored in two earlier (unpublished) pilot studies performed in our laboratory (Remisovsky, Note 1, Note 2). Procedures in both studies involved first obtaining from several independent judges spontaneous verbal descriptions of each of 10 test faces (achromatic 5 x 7 in. photographs of white women). Judges described the faces while viewing the portraits. Facial characteristics mentioned and similarly described (e.g., large eyes, full lips, curly hair, etc.) by three of the five judges were incorporated into a single standard ("objective") description of each face. In the second phase, 32 subjects were shown each of the 10 test faces individually for a brief interval and then were asked to describe the physical features of each face after it was removed from view. Talent for describing faces from memory was measured by comparing subjects' descriptions with objective descriptions. In the last phase of both studies, face recognition memory was tested using completely unfamiliar faces. No relationship between descriptive ability and recognition memory emerged from the results of these studies.

The present study was initiated to once again explore the relation between face recognition memory performance and accuracy of verbal description, using procedures designed to obtain better objective descriptions of test portraits and more accurate measures of subjects' talents for verbally describing faces from memory.

METHOD

General Design

The procedure employed provided subjects differing in relative ability to describe facial features from memory. First, standard descriptions were devised to serve as a criterion for evaluating the accuracy of subjects' verbal portraits. To accomplish this, several judges, while looking at a series of photographs of faces, selected from a checklist of appropriate adjectives those terms that best described characteristics of each of the faces. In this manner, consensual descriptions were developed for each face. Second, the accuracy of these standard verbal descriptions was verified by another set of judges, who tried to match the verbal descriptions with the faces described. Successful matching of faces and descriptions was taken as evidence of the fidelity of the verbal descriptions. Third, immediately after seeing a face, subjects were asked to use the adjective checklist to develop from memory a verbal description for each face. Accuracy of facial description was determined by comparing subjects' descriptions with the standard descriptions. From this comparison, an accuracy score was derived for each subject. In the final phase of the study, all subjects were tested for recognition memory for faces using a different set of faces as stimuli.

Facial Features Adjectives List

A comprehensive list of facial characteristics composed of several commonly seen variants of each feature was developed based partially on the work of Zavala and Paley (1972; see also "portrait parle" of Allison, 1973). This instrument, christened the Facial Features Adjectives List (FFAL) and composed of 23 items (see Table 1), serves as a memory probe by providing two

Copyright © 1979 Psychonomic Society, Inc.

0090-5054/79/010015-04$00.65/0

Table 1
Sample Items from Facial Feature Adjective List

1. Shape of Face–Square, round, oval, V-shaped.
2. Shape of Face–Full, thin, average.
14. Eyebrows–Slanted, straight, arched, curved.
15. Eyebrows–Bushy, thin, average.
23. Shape of Eyes–Almond, round, slitted, average.

to four adjectives commonly used to describe every important feature of the human face. Thus, all judges and subjects, using this standard checklist, were given equal opportunity to recall and describe all aspects of the faces. Employing this instrument, instead of relying on spontaneous reports recalled from memory, reduced the possibility that describing performance merely reflected frequency or rate of verbal output, a trait not necessarily expected to be related to accurate facial description.

Composite Standard Descriptions

Three men and three women judges (recruited from introductory psychology courses), working independently, selected from the FFAL those items that most closely described each of 10 black-and-white test portraits (5 x 7 in.) of college-age white women. Descriptive terms were selected with the portrait in view. Each picture was viewed separately by the judges, and subjects were given unlimited time to make their descriptions.

If three or more judges selected any one item, the item was included in the composite standard description (CSD) of the face. For example, if four judges selected "oval" as the best description of the shape of Face 6, then oval was considered the "correct" shape description for that face. Three or more judges were in agreement on at least 21 checklist items for each of the 10 test portraits.

Verification of the CSD

In Remisovsky's (Note 1, Note 2) investigations, the standard or "objective" verbal description included only those terms used by three or more judges to describe a face, but the similarity between the verbal portrait and the portrait itself was never measured. In the present study, an additional step in the procedure provided empirical information about the closeness of the match between verbal description and face. Five new judges (from introductory psychology courses), in independent sessions with both the test faces and the CSDs simultaneously in view, tried to match five standard descriptions with 5 of the 10 faces. Different sets of five CSDs were given to each judge. Of the 25 possible matches, only 5 were misidentified. We considered this level of relationship between verbal descriptions and facial portraits to represent adequate evidence of the accuracy of the verbal descriptions, and proceeded in the next phase of the study to use the CSDs as prototypes to measure the accuracy with which subjects could describe the test faces from memory.

Derivation of Descriptive Accuracy Scores

In this phase of the study, the 10 test portraits used to develop the standard descriptions were presented for 2.5 sec each to 22 college women subjects. Immediately following each portrait's presentation, subjects used the FFAL to describe the face they had just seen. All subjects were given instructions about use of the FFAL before viewing the first portrait, and they were also told that it was not necessary to make a judgment about all 23 features for every face, but only to select items from the list if they "thought they had some idea about the facial feature."

For each of the 10 portraits, a subject's accuracy of description was determined by comparing the subject's selection of items with the items in the CSD of that face. If this match was perfect, the subject scored 100% on that portrait. Less than perfect matches, by reason of either "inaccurate" choices of adjectives or omissions, were assigned percentage scores that expressed number of correct matches in relation to the number possible for that picture. For example, if the CSD for Portrait 5 included only 21 of the 23 possible facial features, then 21 would be used as the base from which to calculate each subject's percent agreement (or accuracy) score. A subject's overall accuracy score was the mean of her 10 individual percentage scores.

Facial Recognition Memory

One week following the test of facial description, subjects returned to the laboratory to participate in the final condition, face recognition memory performance. Stimuli were 39 black-and-white projected slides of college-age women (senior yearbook photos) selected to minimize nonfacial cues for recognition. None of these faces had been used in the earlier phases of the investigation. During the study session, 10 randomly selected faces were displayed for 2.5 sec each, separated by about a .5-sec interstimulus interval. Subjects were informed beforehand to expect to be tested on their memory of the 10 target faces. During the recognition test, immediately following the study session, subjects viewed the 10 study faces and the 29 other faces from the set of 39 in a predetermined random order that was identical for all subjects. Subjects recorded each response on prepared answer sheets in one of two columns headed "seen before" and "not seen before." Instructions reminded subjects that only 10 targets had been displayed in the first session, and that all 10 would be shown again in this session. Subjects were requested to try to make no more than 10 "seen-before" responses.

RESULTS

Range of performance in the describing task was surprisingly narrow. The best describer, whose descriptions, on the average, were in agreement with the CSD 63% of the time, was only 34% better than the worst describer, whose responses agreed with the CSD, on the average, only 29% of the time. As a result of this relatively narrow range of performance, identical accuracy scores occurred frequently. Notice that the best describer, on the average, agreed with the CSD on only 14 (63%) of the 23 items. This suggests that, even when subjects are prepared to describe faces and do so within a few seconds of seeing a face, their descriptions are impoverished and imperfect in comparison to descriptions made by judges while viewing the faces. Since the average agreement between subjects' descriptions and the CSD was 51%, or about 11-12 of the 23 items on the checklist, the deficit in descriptive adequacy attributable to the effects of reduced exposure (2.5 sec/face) and to memory processes is substantial.[1]

Subjects' descriptive accuracy scores also varied across the 10 test faces. The single lowest accuracy score was 13% (or approximately 3 items), the single highest was 83% (approximately 19 items), and the mean range across all subjects was 35% (8 items). The narrowest range of performance by one subject was 23%; the widest range was 54%. A clearer view of the consistency of subjects' performance in the face description task can be obtained from the results of a split-half (odd vs. even)

product-moment correlation. This analysis indicates that subjects were moderately consistent ($r = +.68$, $p < .01$) in their verbal descriptive performance in response to the 10 test faces. Test faces were not equally "describable," but this source of variability appears to be of relatively minor importance with the set of faces used in this study. For the 10 test faces, mean correct items never exceeded 12 or fell below 9. From these observations, we suggest that subjects do differ reliably in the ability to construct verbal descriptions of faces from memory. In short, "ability to describe faces" appears to be a useful independent variable, although the evidence for this conclusion should be interpreted with caution, primarily because of the moderate size of the reliability coefficient, and the relatively narrow range of performance exhibited by the subjects.

Although other statistical analyses were performed, the relation between descriptive performance and recognition memory performance as measured by both hit rate and d′ were the two major analyses. A Pearson product-moment correlation employing descriptive accuracy scores and number of correct recognitions yielded a value of +.14 ($p > .05$), indicating almost no association between these two measures. A similar analysis, comparing the descriptive accuracy scores with the d′ measure derived from hits and false alarms registered by each subject in the recognition test, yielded a value of +.15, indicating no relationship between the two measures.

Was there even a weak indication that better recognizers are more accurate describers? A mean difference test could not be applied because the large number of identical descriptive accuracy scores located roughly in the middle range (52%) of the distribution prevented clean categorization of the subjects into "inferior" and "superior" groups. Nevertheless, we explored the possibility that the association between recognition memory and describing ability was obscured by the performance of the "middle group" in the following analysis. Significance of mean differences in hit rate and in d′ were computed when the data from subjects with tied scores ($N = 7$) were added to the data of the 7 highest ranking subjects, and also when added to the data of the 8 lowest ranking subjects. In other words, the recognition memory data of the middle group of subjects were included first with the data of the best describers, then with the data of the worst describers, and t tests were performed on the resultant two sets of means. The results of this statistical analysis, faulted as it is by rule infractions, points to a very weak relationship between verbal description and recognition memory. In general, these analyses indicate that the better describers were also slightly better recognizers. In particular, with the middle group of subjects added to the better describers, the mean hit rate (mean = 7.3) of the "superior" group was significantly higher than the mean of the "inferior" group (mean = 5.8) [$t(20) = 2.14$, $p < .05$], and the mean difference in d′ scores (2.13 vs. 1.32) approached significance [$t(20) = 2.02$, $.05 < p < .10$]. With the middle group of subjects added to the worst describers, the difference between both the mean hit rate (7.7 vs. 6.3) and the mean d′ scores (2.41 vs. 1.56) of the "superior" and "inferior" groups approached significance [$t(20) = 1.92$, $p < .10$; $t(20) = 2.05$, $p < .10$].[2]

DISCUSSION

The results of this and the two preliminary investigations offer little evidence for believing the level of ability to verbally describe faces is predictive of the level of ability to recognize faces. Although it is risky to base a conclusion on results in support of the null hypothesis, risk is reduced when the results of several studies all lead to similar conclusions. In two earlier studies, no reliable relationship between face recognition performance and accuracy of describing faces from memory were found (Remisovsky, Note 1, Note 2). Overlap among the procedures and stimuli used in the three studies was minimal. This fact increases both the persuasiveness and the generality of our findings.

Unless "descriptive ability" is an authentic category variable, the findings of this study may be trivial. Thus, if descriptive ability is not a reliable skill, if a subject's future performance cannot be predicted from past performances on similar stimuli, then any relationship between descriptive ability and recognition memory becomes meaningless. If descriptive ability was unstable, the zero-order correlation obtained in this study would merely indicate the lack of correlation between a stable measure and an unstable measure. (In essence, the mean accuracy scores assigned to each subject would be little more than random numbers.) Although the conclusion drawn would remain unchanged (recognition memory performance is not related to describing ability), the underlying reasons for reaching this conclusion would differ when the category variable is or is not reliable. Evidence for reliability of the descriptive accuracy scores obtained in this study indicates that the problem has not been totally resolved, but the ability to describe a face from memory using some kind of memory prompting technique appears to be a moderately stable skill.

There are additional reasons to trust these findings. Intuitively, recognition memory for pictorial material would be expected to far surpass recall of verbally labeled details of pictures. Put in another way, it would have been ridiculous to expect the subjects in Nickerson's (1965), Shepard's (1967), or Standing, Conezio, and Haber's (1970) studies to describe from memory the literally hundreds, even thousands, of pictures they viewed, yet these subjects correctly recognized enormous numbers of pictures they had seen for a few seconds each. In the same vein, when many faces are the stimuli to be remembered, the task of making meaningful different verbal descriptions of each face becomes almost impossible (Chance & Goldstein, 1976). Imagine trying to recall and distinctly describe 20 or 25 faces each seen for 3-4 sec. Yet, without question, at least 65% of these targets would be correctly recognized (Goldstein, 1977). The point of this discussion: both logic and experience suggest that recognition memory for faces and perhaps other kinds of pictures is largely independent of the recall of details of the stimuli.

Research bearing directly on these issues is sparse. In an early study of the factors involved in face recognition ability, Howells (1938) had some of his subjects try to remember details of the faces. Howells reported that subjects with superior recognition scores were relatively poor at remembering details, but he offered no data to support this conclusion. Hall (Note 3), in a more recent investigation, concluded that a (mock)

eyewitness's ability to recognize a suspect in a line-up even may be impaired when the witness talks to a "police artist" and is required to describe the suspect's face in great detail. One interpretation of this finding, offered by Hall, suggests that a witness's own attempts at reconstructing a face from memory affects his final memory picture, and ultimately affects recognition performance. Thus, in an extended effort to correctly portray a particular facial feature, the witness may be affected by his own small modifications and come to accept a final description that is no longer similar to the feature he initially had "in mind." Hall's (Note 3) data point to the existence of a hazy, easily modified memory trace, one that offers little structure from which to develop a clear verbal description.

In a study of people's ability to reconstruct a face from memory using the Photofit system, a forensic tool developed in England and similar to its American counterpart, Identi-Kit, Ellis and his associates (Ellis & Davies, Note 4) examined the relationship between recognition memory for faces and accuracy of Photofit reconstructions. They concluded that the two skills recognition memory and reconstruction ability had little in common. If we can assume the verbal task is analogous to the task of making a Photofit reconstruction, the Ellis and Davies data also can be considered to agree with the results of the present investigation. In an investigation of the role of verbal coding in face recognition memory performance, Chance and Goldstein (1976) instructed two groups of subjects to verbally code target faces in two slightly different ways, while two other groups were asked to either look at the target faces or judge the ages of the faces. It was found that verbal coding, compared to no coding, played only a minor role in recognition memory performance.

REFERENCE NOTES

1. Remisovsky, J. E. *Race, sex, and describing ability as factors in immediate and delayed facial recognition.* Unpublished honors thesis, University of Missouri, Columbia, Missouri, 1976.

2. Remisovsky, J. E. *Sex and judging ability as factors in immediate and delayed facial recognition.* Unpublished study, University of Missouri, Columbia, 1975.

3. Hall, D. F. *Obtaining eyewitness identifications in criminal investigations: Two experiments and some comments on the Zeitgeist in forensic psychology.* Paper presented at the annual meeting of the American Psychology-Law Society, Snowmass, Colorado, 1977.

4. Ellis, H. D., & Davies, G. M. *An investigation of the photofit system for recalling faces.* Final report, Grant HR3123/1, 1974-1976, Social Science Research Council (Report 8). Department of Psychology, University of Aberdeen, Aberdeen, Scotland.

REFERENCES

ALLISON, H. C. *Personal identification.* Springfield, Ill: Charles C. Thomas, 1973.

CHANCE, J., & GOLDSTEIN, A. G. Recognition of faces and verbal labels. *Bulletin of the Psychonomic Society*, 1976, **7**, 384-386.

GOLDSTEIN, A. G. The fallibility of the eyewitness: Psychological evidence. In B. D. Sales (Ed.), *Psychology in the legal process.* New York: Spectrum, 1977.

HOWELLS, T. H. A study of the ability to recognize faces. *Journal of Abnormal and Social Psychology*, 1938, **33**, 124-127.

NICKERSON, R. S. Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology*, 1965, **19**, 155-160.

SHEPARD, R. N. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 156-163.

STANDING, L., CONEZIO, J., & HABER, R. N. Perception and memory for pictures: Single-trial learning of 2,500 visual stimuli. *Psychonomic Science*, 1970, **19**, 73-74.

ZAVALA, R. T., & PALEY, J. J. (Eds.). *Personal appearance identification.* Springfield, Ill: Charles C. Thomas, 1972.

NOTES

1. When judges were asked to simply describe faces while viewing the portrait, as in Remisovsky's (Note 1) study, no more than six concordant items per picture were generated. In the present study, using a "cued description" technique (the FFAL), at least 21 concordant items were selected by the judges for each of the test faces. Similarly, Remisovsky's best describers (i.e., subjects, not judges) averaged about three concordant responses per portrait; our subjects, using the FFAL, averaged 12 items. Quite clearly, the use of an adjective checklist appreciably improves the accuracy of facial descriptions.

2. In this study, for each subject's performance on each of the 10 test portraits, a descriptive accuracy score was assigned using a formula [(number correct items)/(number CSD items)] that took into account only correct items; both omitted items and incorrect items were considered to be errors. The data have also been analyzed using the formula, (number correct items)/(number items selected by subject). For example, in response to a test portrait, a subject could use only 12 of the 23 FFAL items, but if 10 of these items agreed with the CSD, accuracy would be .83 (or 10/12), instead of .43 (10/23). Either method of analysis leads to the same conclusions.

(Received for publication September 15, 1978.)