Describing Faces from Memory: Accuracy and Effects on Subsequent Recognition Performance

Michael S. Wogalter
Department of Psychology
North Carolina State University
Raleigh, NC 27695-7801

ABSTRACT

The present research examines: (a) the accuracy of three face description methods, and (b) the effects of post-exposure description and imaging activities on subsequent face recognition performance. Participants viewed a sequence of six target photographs, and after each, performed one of three description tasks: generated their own set of descriptors, checked-off descriptors from a pre-existing list, or rated the same set of descriptors on bipolar scales. Other participants performed a distractor (control) activity. Additionally, participants were either told or not told to image the targets while they simultaneously performed the description tasks. Results showed that the checklist task lowered subsequent recognition performance compared to the generate task. Imaging with the generate task facilitated recognition, but imaging with the checklist and rating tasks degraded recognition. The generate task produced the highest quality descriptions as determined by other participants' performance in matching the descriptions to face photographs. The checklist decrement is discussed in terms of memorial confusion initiated by the presence of irrelevant face cues. These results indicate that descriptor generation is the preferred method of collecting eyewitness' face descriptions.

INTRODUCTION

Eyewitnesses, having viewed an individual involved in a crime, are frequently asked by police investigators to participate in various memory tests, e.g., examining a lineup or a mugfile. In the intervening period prior to these tests, several things might happen: witnesses may rehearse or hold an image of the face in their mind and/or they might participate in other memory tests such as giving a verbal description or helping to produce a sketch or composite.

The present research examines three types of verbal description instruments with respect to the quality of descriptions that they produce, and whether these verbal description methods (and holding an image of the target face) affects subsequent recognition performance.

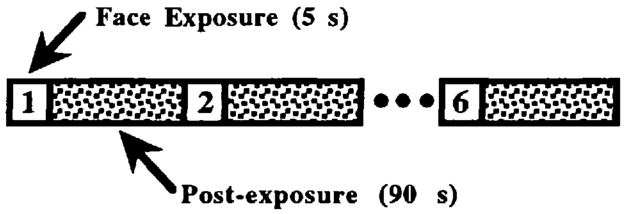
Why would different methods of eliciting verbal descriptions be of interest? Witnesses generally do not give many descriptors when describing faces, and the terms that they produce are frequently not very specific (e.g., 'medium nose' or 'thin lips'). These general terms are frequently inadequate to do an effective search for the culprit. One possible way to elicit better descriptions is to provide a list of possible adjective descriptors which they can select from—as opposed to generating the descriptors themselves.

There has not been much empirical work directly comparing different description techniques. One exception is by Goulding (1971) who had police officers make cued or free recall descriptions of target faces. Free recall produced better quality descriptions than cued recall. In the present study, this effect is re-examined and includes ratings as another kind of cued method.

Another issue is whether the description task influences subsequent recognition. Previous research on this topic is equivocal. Verbal description has been reported to degrade, facilitate, and have no effect on subsequent recognition performance. However, virtually all of the previous studies used different methods of eliciting the descriptions. The exception is Wogalter (1991) who compared two description methods, and found that a cued test which made available specific descriptors (checklist) produced lower subsequent recognition compared to a description test in which participants produced their own terms (generate). One possible reason for this decrement is that the checklist, by its very nature, has numerous descriptors that are irrelevant, or wrong, with respect to any particular target face. By considering these terms, witnesses might get confused on what the target looked like, lowering subsequent recognition. In the present study, this issue is re-examined using a different checklist and generate test, plus adds a third method involving ratings.

Another activity the witness might perform is to visually rehearse or image the target face. Some studies (Graefe and Watkins, 1980; Read, 1979) have shown a *small* facilitative effect on recognition. However, other research has found no effect (Schooler and Engstler-Schooler, 1990) or a negative effect (Hall, 1979) of imaging. The influence of imaging in the present study is examined in a way heretofore not examined: Imaging is concurrent with the assigned post-exposure activity.

Figure 1. *Exposure/Post-exposure sequence for the 6 targets.*



METHOD

Design and Participants

The experiment was a 4 (post-exposure task) X 2 (imaging instructions) between-subjects design A total of 192 undergraduates participated in the main experiment, 24 in each cell. Another 12 and 4 students participated in the in-view description and matching tasks, respectively.

Procedure and Materials

Six white male targets (in frontal poses) were selected at random from a large pool of photographs. Targets were shown in the initial exposure (study) as color slides for 5 s each. Following each target slide, a 90 s period was provided where participants were to perform one of the post-exposure activities described below. Figure 1 displays a representation of the exposure/post-exposure sequence.

Participants were given a booklet containing six pages that differed according to post-exposure condition. Representations of the pages in the three description conditions can be seen in Figure 2.

(1) In the *Rate* condition, the pages of a response booklet contained a list of 10 feature headings. Under each heading was a set of 5-point bipolar scales with adjective descriptor endpoints reflecting various dimensions. The descriptors and dimensions were derived from previous verbal description research, and studies on feature saliency, multidimensional scaling and cluster analyses of faces. The

Figure 2. *Representations of the Post-exposure description forms.*

Rate

OVERALL SHAPE OF FACE		
short	|_|_|_|_|_|	long
narrow	|_|_|_|_|_|	broad
bony	|_|_|_|_|_|	fleshy
COMPLEXION		
fair	|_|_|_|_|_|	dark
pale	|_|_|_|_|_|	red
unlined	|_|_|_|_|_|	lined
clear	|_|_|_|_|_|	blemished
HAIR		
short	|_|_|_|_|_|	long
tidy	|_|_|_|_|_|	untidy
straight	|_|_|_|_|_|	curly
bald	|_|_|_|_|_|	thick/full
black	|_|_|_|_|_|	white
FOREHEAD		
low	|_|_|_|_|_|	high
narrow	|_|_|_|_|_|	broad
straight	|_|_|_|_|_|	sloping
EYEBROWS		
thin	|_|_|_|_|_|	thick
straight	|_|_|_|_|_|	bent
low	|_|_|_|_|_|	high
meets in middle	|_|_|_|_|_|	set far apart
EYES		
small	|_|_|_|_|_|	large
narrowed	|_|_|_|_|_|	opens
close set	|_|_|_|_|_|	wide spaced
deep set	|_|_|_|_|_|	protruding
dark	|_|_|_|_|_|	light
NOSE		
small	|_|_|_|_|_|	large
short	|_|_|_|_|_|	long
narrow	|_|_|_|_|_|	broad
concave	|_|_|_|_|_|	hooked
small nostrils	|_|_|_|_|_|	large nostrils
narrow tip	|_|_|_|_|_|	broad tip
MOUTH		
small	|_|_|_|_|_|	large
thin upper lip	|_|_|_|_|_|	thick upper lip
thin lower lip	|_|_|_|_|_|	thick lower lip
CHIN		
small	|_|_|_|_|_|	large
pointed	|_|_|_|_|_|	square
receding	|_|_|_|_|_|	jutting

Checklist

OVERALL SHAPE OF FACE			
bony	_____	narrow	_____
short	_____	long	_____
fleshy	_____	broad	_____
COMPLEXION			
pale	_____	dark	_____
lined	_____	clear	_____
fair	_____	red	_____
unlined	_____	blemished	_____
HAIR			
bald	_____	white	_____
untidy	_____	short	_____
thick/full	_____	black	_____
tidy	_____	curly	_____
straight	_____	long	_____
FOREHEAD			
low	_____	high	_____
narrow	_____	broad	_____
straight	_____	sloping	_____
EYEBROWS			
straight	_____	low	_____
set far apart	_____	bent	_____
thin	_____	high	_____
meets in middle	_____	thick	_____
EYES			
open	_____	dark	_____
mall	_____	protruding	_____
close set	_____	large	_____
light	_____	narrowed	_____
wide spaced	_____	deep set	_____
NOSE			
narrow	_____	large	_____
hooked	_____	short	_____
small nostrils	_____	broad	_____
broad tip	_____	concave	_____
small	_____	narrow tip	_____
large nostrils	_____	long	_____
MOUTH			
thin upper lip	_____	thick lower lip	_____
small	_____	thin lower lip	_____
large	_____	thick upper lip	_____
CHIN			
jutting	_____	small	_____
large	_____	square	_____
pointed	_____	receding	_____

Generate

OVERALL SHAPE OF FACE

COMPLEXION

HAIR

FOREHEAD

EYEBROWS

EYES

NOSE

MOUTH

CHIN

descriptors/dimensions have been used successfully in the FRAME computer-assisted search system to locate target faces in a mugfile (Shepherd, 1986). Participants were told to complete the form to describe the face just seen. (2) In the *Checklist* condition, the sheets were identical except the adjectives under each heading was randomized. Participants were to check all descriptors that described the previously seen face. (3) In the *Generate* condition, the sheets were identical to the other conditions except the adjective descriptors were deleted. This provided space for participants to write descriptions in their own words. (4) Participants in the *Distractor/Control* condition were told that a second purpose of the study was to measure how fast they could do a visual scanning/perceptual speed task. The pages of the booklet contained a large matrix of random letters. Two different letters were circled on each page. Participants were told that during the post-exposure periods they should mark all other instances of the circled letters on the page. Performance on the control task was not analyzed.

Half of the participants received explicit instructions to generate and hold a mental image of the most recently viewed face while they simultaneously worked on one of the post-exposure tasks. The other half of the participants were not given explicit instructions to image.

Figure 3. *Recognition Performance Measures.*

Recognition Responses	***Confidence Rating***
N = No, not presented	1 = guessed
Y = Yes, presented	2 = probably correct
	3 = certain

6-Point Scale that combines the above two measures

"No" responses			"Yes" responses		
"N3"	"N2"	"N1"	"Y1"	"Y2"	"Y3"
1	2	3	4	5	6

6 Recognition Measures:

- **Hit scores (for 6 Targets):**
  - (a) **HM:** Hit/miss (Mean of targets on 6-point scale)
  - (b) **PH:** Proportion hit (Mean of targets where "Y" = 1, "N" = 0)
- **False alarms scores (for 134 Distractors):**
  - (c) **FACR:** False alarm/correction rejection (Mean of distractors on 6-point scale)
  - (d) **PFA:** Proportion false alarm (Mean of distractors where "Y" = 1, "N" = 0)
- **Discrimination/Sensitivity scores:**
  - (e) **H-F:** HM minus FACR
  - (f) **SHM:** Mean target z-score (after standardizing each participant's responses to all test photographs)

Note: Better recognition performance is indicated by ***higher*** scores on the ***hit*** and ***discrimination*** measures and ***lower*** scores on the ***false-alarm*** measures.

After the exposure/post-exposure phase was completed, participants worked on a study strategy questionnaire for 5 min which was immediately followed by the recognition test. The test pictures were comprised of 140 black and white slides: 134 were distractor faces, and the other 6 were the targets. The targets appeared at random positions after the 50th distractor slide. The target photos in the test sequence were similar, but not identical, to those of the shown at study; they were taken by different cameras several minutes apart. Participants were told that the faces they saw at study may or may not be in the test sequence.

In the recognition test, slides were presented at a 7 s rate. Participants indicated "yes" or "no" to each face according to whether they believed the individual was shown earlier, and also gave a 3-point confidence rating. From these scores, 6 recognition performance measures (two hit, two false-alarm, and two discrimination) were derived as shown in Figure 3.

RESULTS

Recognition Performance

Recognition performance was examined using 4 (post-exposure task) X 2 (imaging instructions) between-subjects analyses of variance (ANOVAs). Comparisons among means for significant effects were performed using Fisher's Least Significant Difference test. The comparisons described below are at or below the .05 probability level.

The hit measure means, HM and PH, are shown on the top row of each cell in Table 1. The ANOVAs showed a significant effect of post-exposure task, $F(3, 184) = 13.84$, $MS_e = .672, p < .0001$, and $F(3, 184) = 11.38$, $MS_e = .033, p < .0001$, respectively. The Distractor condition produced significantly lower HM and PH scores compared to each of the three verbal description conditions. Neither hit measure showed a main effect of imaging instructions, $Fs < 1.0$. However, imaging instructions interacted with post-exposure task, $F(3, 184) = 4.25$, $MS_e = .672$, $p < .01$, and $F(3, 184) = 3.58$, $MS_e = .033$, $p < .02$, for HM and PH, respectively. With HM, the Checklist and Rate conditions showed a significant decrement with imaging instructions. With PH, Rating plus imaging produced a decrement, but Generating plus imaging produced facilitation.

The second row of each cell in Table 1 shows the mean false alarm scores. There was a significant main effect of post-exposure task with FACR, $F(3, 184) = 11.38$, $MS_e = .033$, $p < .0001$, but not with PFA. The Checklist produced significantly higher FACR than the Generate or Distractor conditions. Neither false alarm measure yielded a main effect of the imaging instructions or interaction ($Fs < 1.0$).

The two discrimination scores, H-F and SHM, are shown on the bottom row of each cell in Table 1. The

ANOVAs showed significant post-exposure main effects with both measures, $F(3, 184) = 10.25, MS_e = .770, p < .0001$, and $F(3, 184) = 4.94, MS_e = .553, p < .0001$ for H-F and SHM, respectively. The Generate condition produced significantly higher discrimination than the Checklist condition. The Distractor condition produced significantly lower discrimination than the three description conditions (except the Checklist with SHM). Neither discrimination measure showed a main effect of imaging instructions ($Fs < 1.0$), but there was a significant interaction with H-F, $F(3, 184) = 2.80, MS_e = .770, p < .05$, that was marginal for SHM, $F(3, 184) = 1.84, MS_e = .553, p < .07$. The Checklist produced significantly lower H-F with imaging instructions than with no imaging instructions. With SHM, the Generate condition tended to be higher with than without imaging ($p = .06$).

Additional analysis examining performance for only the first face that was shown to participants indicated a pattern of results that was virtually identical to those described above. Other analyses showed that target face did not interact with post-exposure task or imaging condition.

Description Quality

To examine description quality, a separate group of participants completed the description forms while the targets were in view (i.e., not from memory). These in-view descriptions and all of the descriptions produced in the main (post-exposure) experiment were randomized, assembled into booklets, and then four participant judges attempted to match the descriptions to the six target photographs mounted on a poster board. From the matching assignments, a measure of quality was derived.

If the descriptions did not provide any useful information, the judges would make their matching assignments at random and performance would be at or near the chance level of 1/6 or .167. Table 2 shows the means. Performance was better than chance for all conditions ($ps < .001$) indicating that the description techniques provide, at least, some useful information.

Post-exposure description quality. A 3 (post-exposure description task) X 2 (image instructions) between-subjects ANOVA using the description quality measure showed a significant effect of description task, $F(2, 138) = 36.25, MS_e = .373, p < .0001$. Paired comparisons indicated that the Generate condition produced higher quality descriptions than the Checklist and Rate conditions. The ANOVA showed no main effect of imaging instructions ($F < 1.0$), but this factor interacted with description condition, $F(2, 138) = 5.48, MS_e = .373, p < .01$. Only one pairwise comparison was significant: Higher quality Checklist descriptions were produced with imaging instructions than without imaging instructions.

In-view description quality. The in-view description quality means are shown on the bottom row of Table 1. A one-way within-subjects ANOVA showed a significant effect of description task, $F(2, 22) = 6.00, MS_e = .246, p < .01$. Pairwise comparisons indicated that the Generate condition produced significantly better descriptions than the Checklist and Rate conditions.

Relation of Description Quality and Recognition

Correlational analyses examined the relation between post-exposure description quality and recognition performance. For the Checklist, none of the recognition measures was significantly related to description quality. However, for the other two description conditions, there were significant positive correlations between quality and recognition discrimination (Rate: H-F, $r = .38$, $n = 48$, $p < .01$, SHM, $r = .35$, $n = 48$, $p < .05$; and Generate: H-F, $r = .35$, $n = 48$, $p < .05$, SHM, $r = .38$, $n = 48$, $p < .01$).

Table 1. Mean Recognition Performance as a Function of Post-Exposure Description and Imaging Condition.

		Post-Exposure Tasks							
		Checklist		Rate		Generate		Distractor/Control	
Image Instructions	HM (PH)	4.71	(.81)	4.56	(.76)	5.13	(.89)	4.24	(.69)
	FACR (PFA)	2.45	(.25)	2.09	(.20)	2.08	(.19)	2.14	(.21)
	H-F (SHM)	2.26	(1.50)	2.47	(1.74)	3.05	(2.14)	2.10	(1.45)
No Image Instructions	HM (PH)	5.19	(.88)	5.11	(.88)	4.85	(.81)	3.87	(.63)
	FACR (PFA)	2.45	(.27)	2.35	(.25)	2.12	(.20)	2.12	(.20)
	H-F (SHM)	2.74	(1.78)	2.76	(1.85)	2.73	(1.75)	1.75	(1.32)
Mean	HM (PH)	4.95	(.84)	4.84	(.82)	4.99	(.85)	4.06	(.66)
	FACR (PFA)	2.45	(.26)	2.22	(.22)	2.10	(.20)	2.13	(.20)
	H-F (SHM)	2.50	(1.64)	2.62	(1.80)	2.89	(1.95)	1.93	(1.39)

Table 2. Mean Post-Exposure and In-View Description Quality as a Function of Description and Imaging Conditions.

	Description Method		
	Checklist	*Rate*	*Generate*
Imaging Instructions	.46	.43	.62
No Imaging Instructions	.32	.46	.67
Mean	.39	.44	.64
In-view	.65	.60	.77

DISCUSSION

The results show that recognition performance following the Checklist task was lower than the Generate task. This result may seem somewhat surprising given the fact that participants only had to check off appropriate descriptors. Lowered performance in the Checklist condition can be explained in terms of exposure to irrelevant or wrong descriptors. By its nature, the checklist provided extraneous descriptors (in order to describe a range of different faces). Therefore, some adjectives were not descriptive of the particular face they had just viewed. By considering these erroneous terms, participants possibly incorporated some of this information into memory, resulting in confusion about what the target looked like, reducing subsequent recognition performance. The Rate technique provided the same descriptors as the Checklist but produced a less severe decrement. In the rating task the descriptors were ordered along dimensions which might have enabled consideration of a broader range of features, thereby causing less confusion. In contrast, the Generate technique allowed participants to produce verbal descriptions without the confusion of irrelevant descriptors because the terms were not present. Nevertheless, Schooler and Engstler-Schooler (1990) reported recognition interference using a generate-type description task following exposure. However, these researchers provided a much longer period of time to describe the face (5 min) which could promote confabulation of irrelevant face features while composing the description. The theorized confusion of memory by intervening stimuli is similar to the interference reported in other research (e.g., Loftus and Greene, 1980) and supports earlier work (Wogalter, 1991) showing a recognition decrement with a different descriptor checklist.

In this experiment, strong support for an overall benefit of imaging on recognition was not found. However, imaging instruction interacted with the verbal description tasks showing some improvement when it co-occurred with the Generate method, and a decrement when it co-occurred with the Checklist and Rate methods. One explanation for these results is that the request to image in the Checklist and Rate conditions increases the likelihood that participants imaged representations of irrelevant (or wrong) verbal descriptors But when directed to image in the Generate condition, participants could do so without irrelevant terms to consider.

The description quality results showed that all three description techniques provide some useful descriptive information, but the Generate condition produced the best descriptions, under both post-exposure and in-view conditions. The Generate technique allows greater freedom to use the most effective language to describe the targets. The other two description tasks are more restrictive in the features that could be described.

The results also showed that directing participants to image produces significantly better quality Checklist descriptions than without these instructions. One explanation is that Checklist participants, without explicit instructions to image, might merely check off descriptors with less considered thought than those given image instructions. However, the process of imaging irrelevant items on the Checklist might partially destroy specific target memory — producing degraded recognition in the subsequent test.

Neisser (1987) suggests that free recall tests are more accurate and less likely to produce distorted, constrained, contaminated memorial reports than cued recall and recognition tests. The present results support this notion. When capturing face descriptions, free recall methods (like the generate condition) are preferred over methods that rely on recognition of descriptors (like checklists), because descriptor generation does not degrade subsequent recognition, and it produces the better quality descriptions.

REFERENCES

Goulding, G. J. (1971). Facial description ability. *Police Research Bulletin, 19*, 42-44.

Graefe, T. M., and Watkins, M. J. (1980). Picture Rehearsal: An effect of selectively attending to pictures no longer in view. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 156-162.

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

Loftus, E. F. and Greene, E. (1980). Warning: Even memory for faces may be contagious. *Law and Human Behavior, 4*, 323-334.

Mauldin, M. A., and Laughery, K. R. (1981). Composite production effects on subsequent facial recognition. *Journal of Applied Psychology, 66*, 351-357.

Neisser, U. (1987). *The present and the past*. Paper presented at the Second International Conference on Practical Aspects of Memory, Swansea, Wales, U.K.

Read, J. D. (1979). Rehearsal and recognition of human faces. *American Journal of Psychology, 92*, 71-85.

Schooler, J. W., and Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36-71

Shepherd, J. W. (1986). A computer-based face image retrieval system. In H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. W. Young (Eds.), *Aspects of face processing*. Dordrect: Martinus Nijhoff.

Wogalter, M. S. (1991). Effects of post-exposure description and imaging on subsequent face recognition performance. *Proceedings of the Human Factors Society, 35*, 575-579.