Verbal Descriptions of Faces From Memory: Are They Diagnostic of Identification Accuracy?

Gary L. Wells
University of Alberta, Edmonton, Canada

There is considerable forensic import to the hypothesis that the quality of an eyewitness' description of a face is useful for predicting the accuracy of a subsequent identification. Although previous studies have failed to support this hypothesis, those studies were designed to test only whether witnesses who are good describers also are good identifiers. The current study used 88 different target faces and found a significant point-biserial correlation between description accuracy and identification accuracy. This relationship was not due to a process wherein good describers are good identifiers, however, but was due to the fact that faces that are better described are better identified; a relationship that could not be tested in the designs of previous studies. The quality of a subject's description of a given face did no better than did a second subject's description of that face in terms of predicting the former subject's identification accuracy. Because the description-identification relationship was found to be mediated by target factors (rather than subject characteristics), it was suggested that different assessments (e.g., of target-face uniqueness) could better predict identification accuracy.

The U.S. Supreme Court has endorsed explicitly the belief that there is a meaningful and useful relationship between an eyewitness' verbal description of a perpetrator and the accuracy of a subsequent lineup or photographic identification (Neil v. Biggers, 1972). A recent review of the experimental literature, however, calls this belief into question (Wells & Murray, 1983). Specifically, research seems to indicate that people who are superior at describing details of faces from memory are not appreciably superior at recognizing faces (e.g., Goldstein, Johnson, & Chance, 1979; Howells, 1938; Wolfskiel & Brigham, 1985).

There are numerous reasons why verbal descriptions of faces and recognition of faces might have little or no relationship. First, research suggests that verbal processes play little or no role in the encoding of faces (Ellis, 1984). It may be that language is too imprecise to be of much use in processing faces. Nevertheless, there are aspects of faces that can be described directly or from memory (e.g., hair color, facial hair, nose size, scars, eye color) and these would appear to have at least some potential diagnostic value. Another reason why verbal descriptions and identifications may be unrelated is that the cues to retrieval for these two tasks may be uncorrelated. Verbal description is a recall task in which context is provided to some degree (e.g., via orienting instructions to think back to the scene of the crime) and the individual must retrieve the target information. Identification, however, is a recognition task in which the target item (or a similar but incorrect alternative item) is provided and the individual must retrieve contextual aspects of the original episode (was this the person who . . . ?). Research indicates that retrieval cues in these two situations may be uncorrelated (e.g., Broadbent & Broadbent, 1977; Flexser & Tulving, 1978; Tulving & Watkins, 1977). Indeed, this distinction between recall and recognition helps account for the finding that whether or not a word is recalled has no predictive value for whether or not it will be recognized (Flexser & Tulving, 1978).

Studies that have come closest to testing the Court's (Neil v. Biggers, 1972) hypothesis are those of Goldstein, Johnson, and Chance (1979) and Wolfskiel and Brigham (in press). Unfortunately, Goldstein et al. did not ac-
tually test whether or not an eyewitness' description of a face relates to the eyewitness' identification accuracy for the face. Instead, their research asked whether people who are "good" describers are also good identifiers. Specifically, their subjects were shown faces which they had to describe from memory and were later shown a different set of faces that they had to recognize from memory. In other words, the faces subjects described were not the ones they had to recognize. Goldstein et al., therefore, did not test the Court's hypothesis, nor did they intend to test that hypothesis. Nevertheless, their research is often cited on the issue and it is not totally irrelevant. Their research suggests that one aspect of the relationship between description and identification accuracy is absent (i.e., there appear to be no stable individual differences that relate identification accuracy and quality of description).

Wolfskiel and Brigham (in press) conducted the best test of the hypothesis to date. They exposed subjects to one of two target persons under conditions in which subjects made deep or shallow judgments of the target at the time of encoding. Later, subjects gave verbal descriptions and then attempted to identify the target person among distractors. Their results indicated that subjects who gave relatively accurate descriptions were not more likely than those who gave relatively poor descriptions to identify the actual target. Although the Wolfskiel and Brigham (in press) study is a good test of the Court's hypothesis at one level there remains another level at which the court's hypothesis may prove to have some validity. Specifically, the Wolfskiel and Brigham study represents a within target-face analysis of the hypothesis, whereas the current study investigates the description-identification relationship across target faces. The distinction is critical at a conceptual level and could produce different conclusions. Consider, for example, a study that exposes subjects to a target face that they later describe and try to identify. Suppose, however, that the target face is the same for all subjects. Any relationship between description accuracy and identification accuracy that emerges from such a design would have to result from a process wherein subjects who are good at describing the target's face are also good at identifying the target's face.

Suppose, however, that there is a natural correlation among faces themselves wherein some faces are both easy to describe and easy to identify while other faces are both difficult to describe and difficult to identify. Indeed, there is good evidence that some faces (unique ones) are much easier to identify than other faces (Goldstein & Chance, 1981), and it seems reasonable to assume that these same (unique) faces may also be easier to describe. If this is true, then it should also be true that the "goodness" of an eyewitness' description of a perpetrator has some predictive relationship to the likelihood of a correct identification of the perpetrator. This particular source of relationship between description and identification cannot be tested with a design that uses only one target face. Similarly, a design that uses only two target faces (as did Wolfskiel & Brigham, in press) cannot test the hypothesis because the source of correlation is hypothesized to reside in the target faces themselves. Therefore, the current study used a large number of target faces in which the target faces were treated as a random variable.

Whenever a random variable is used there is some issue about the extent to which it is representative of the distribution under consideration. Because the hypothesis concerns the Courts Bigger's criteria, a strict interpretation of random would require a sample of faces of criminal suspects. No claim is made in this study for such an extreme interpretation of representativeness. The faces used in this study are representative only of the three large undergraduate classes from which target faces were selected. Nevertheless, if the hypothesized relationship occurs with the current sample of faces, there seems no good reason to assume that it would not also occur in a sample of faces of criminal suspects.

One other issue deserves mention with regard to the forensic significance of the description-identification hypothesis. Although Court used the phrase "accuracy of description," the Court actually meant "agreement of description" or "congruence of description." Accuracy is the wrong term because there is no way in an actual case to judge the accuracy of an eyewitness' description unless one assumes the defendant to be guilty. What the court meant was that the congruence between a prior verbal description of the perpetrator and the physical characteristics
of the identified person has probative value in judge the likelihood of the identification being accurate. Thus, to test the Court's hypothesis, we must measure the relationship between the verbal description and the physical characteristics of the identified person and see if that relationship in turn predicts the accuracy of the identification. In other words, if prior descriptions by witnesses are to have any predictive utility in actual cases, it will not be because the description can be compared to the actual perpetrator (it cannot in actual cases) but because the description can be compared to the identified person (who may or may not be the actual perpetrator). Therefore, both description congruence and description accuracy measures are used in this study.

Overview

Each of 176 subjects viewed one of 88 faces in one of two profiles (straight on versus three-quarters) and made trait judgments of the face. Following an interpolated task, subjects were asked to describe the face. Following a second interpolated task, subjects were required to identify the face among a set of 21 photographs that either did or did not contain the original face. A face viewed initially in one profile was tested for identification with the other profile. Two judges scored the descriptions for accuracy and congruence and these description scores were then related to the dichotomous identification accuracy measure through the use of point-biserial correlations. Because each face was viewed by 2 subjects, it was also possible to examine the relationship between one subject's description of a given face and another subject's identification of that same face, yielding an estimate of the extent to which the faces themselves are a source of any description-identification relationship.

Method

Materials

Target faces were obtained by photographing 230 caucasian persons (139 female, 91 male) ranging in age from 17 to 33 years. Each person was photographed in both a three-quarters profile and a straight-on pose. Target persons wore high-collar lab coats to hide any clothing cues to recognition. Because the hypothesis states that between-face variance might mediate a description-identification relationship, efforts were taken to assure that the faces were not excessively homogeneous. This is made difficult by the possibility that attractive persons, or persons who perceive themselves as attractive, are more likely than less attractive persons to agree to be photographed. This bias was minimized, however, by persuading intact university classes to be photographed. The use of both an incentive ($3.00) and a persuasive communication regarding the importance of getting variance in face types was successful in getting 100% compliance in the three classrooms (N= 106, 82, 42). These photographs were taken two and one-half years prior to the participation of the subjects in the current experiment, thereby minimizing chances that the subjects knew the target persons.

The sample of target faces were, of course, restricted in terms of age. However, the forensic logic of the study is not harmed appreciably by this restriction because the hypothesis states merely that faces that are easy (difficult) to describe may be the ones that are easy (difficult) to identify. Thus, although it is important to represent population variance among target faces on the ease of description, there is no strong reason to believe that 20 year olds, for example, show more or less variance than 50 year olds. Admittedly, however, the age restriction (as well as the fact that the targets are all university students) would be problematic for null results.

Forty-four photographs of each sex were selected by deleting randomly 95 female photographs and 47 male photographs. Within each set of 44 photographs, a separate sample of 15 subjects viewed each photograph with one other photograph with which it was judged most similar. The purpose of this was to create a plausible "foil" for each target such that whenever the identification set did not contain the original target, the foil was substituted. The forensic logic of having target-absent identification conditions is discussed more fully elsewhere (e.g., Lindsay & Wells, 1980; Malpass & Devine, 1981; Wells & Lindsay, 1980). Briefly, it is necessitated by the fact that mugshots, lineups, and photo-lineups might not include the actual perpetrator in real-world cases. Each pair of photographs within each group of 44 was then assigned randomly to one of two replication groups, yielding 4 sets of 22 photographs (2 sets of each sex).

Eight photo albums were used to display the photographs for purposes of the identification task. Four of the albums held straight-on pose photographs and four held the three-quarters profile photographs. The purpose of the different profiles was to assure that the identification task was a profile identification task rather than a photograph identification task. That is, any peculiar characteristics of a photograph (e.g., a temporary light intrusion) might serve as an ecologically invalid cue to retrieval, thereby making the task one of photo identification rather than identifying the person. Thus, subjects who were exposed to a three-quarters profile at encoding were tested with a straight-on pose at retrieval and vice versa. Each identification album contained 21 photographs rather than 22 because of the presence/absence of the original target. That is, if a subject was shown person number 17 in the original task, that subject was assigned to view an identification album that either contained the original target person or the target's substitute. When the original target was in the identification album the substitute was not and vice versa. Each of 7 pages of the album...
Subjects, Design, and Procedure

One-hundred and eighty undergraduate psychology students participated in partial fulfillment of a course requirement. Subjects were assigned randomly to view one of the 88 possible target persons either in straight-on pose or three-quarters pose. Four subjects indicated that they previously had seen or knew personally the target person and their data were eliminated. Thus, two subjects viewed each target person one under each profile condition. Subjects were tested individually and were told that the study concerned inferences that people make about other people. Subjects were then shown a target photograph displayed in a photograph holder and were given a 10-item adjective rating form which asked them to judge the person on characteristics of honesty, intelligence, friendliness, aggressiveness, kindness, stinginess, imaginativeness, arrogance, selfishness, and excitability. Ratings were on 7-point scales. Previous research (e.g., Bower & Karlin, 1974) shows that such tasks promote incidental encoding of faces in memory. Subjects were told that they could have 4 min to do the task; all subjects completed the task within the allotted time.

Subjects were then given a distractor task in which they filled out a questionnaire on research participation, describing their experiences and opinions. The questionnaire was 7 pages long and was given ostensibly to fill in the time that the experimenter needed to prepare for phase two of the experiment and to help assess need for improvement in the research participation procedures used by the department. After 10 min at this questionnaire, which no subject had finished, the experimenter asked the subject to put it aside until later so as to go on to the next phase of the study. At that point the experimenter gave the subject a form that asked for a description of the target person’s photograph that they were shown earlier. The form said:

Please try to describe the person whose photograph you were shown. Specifically, your task is to describe the person in such a way that your description would aid someone else in attempting to identify the person. Your description should, therefore, focus primarily on physical features. You might begin, for example, by describing the person’s sex, race, age, hair color and style, facial features and so on. Try to be as complete as possible but do not guess at things for which you are uncertain. If you need more space than that provided below, simply turn over the page and continue on the other side. Print clearly.

Subjects were told that they had up to 12 min to finish their description and the experimenter left the room. The experimenter returned after 12 min. All had finished within the allotted time. The experimenter then picked up the descriptions and gave subjects a second filler task. This task required subjects to make probability judgments for various events and subjects were told that they had up to 10 min to complete as many as possible. There were 106 such events and no subject had completed this filler task when the experimenter returned 10 min later. At that point the experimenter retrieved the filler task materials and placed one of the eight photo albums in front of subject in closed position. The experimenter explained that the subject was now to attempt to identify the person who they had seen earlier and that the person might or might not be in the set of 21 photographs contained in the photo album. An identification form was given to the subject on which to indicate his or her choice. Each photograph was associated with a number in the identification album. Subjects were asked to open the album and the experimenter exited immediately. The experimenter returned after 5 min, collected the identification album and identification response form, debriefed the subject, and answered any questions.

Results

The 176 descriptions were each scored independently by two judges and each judge scored each description in three ways. First, descriptions were scored for completeness by counting the number of facial features mentioned in the description. The list of features from which the judges worked included hair color, hair style, nose, lips, eyes, brows, facial hair, chin, cheekbones, ears, and any other distinguishing marks (e.g., scars). None of the targets wore eyeglasses. Descriptions taken from a previous study were used to train the judges. Disagreements occurred occasionally. Almost always these were due to one judge considering a description such as “his hair was long and straight” to be two features, whereas the other considered it to be one feature. In such cases, the average (i.e., 1.5) was used. Descriptions were then scored for accuracy by having the judges assign a value of −1, 0 (zero), or +1 to each feature mentioned in the description according to whether it was judged inaccurate, ambiguous, or accurate, respectively with regard to the original target person. Judges had both poses with which to compare the descriptions. Again, an average was used in cases of disagreement between the judges. Finally, descriptions were scored for congruence by having judges assign a value of −1, 0 (zero), or +1 for each feature (as with the accuracy measure) except that the comparison was between the description and the chosen photograph. For 40% (71) of the subjects their choice was accurate and another 28% (49) made no choice. Therefore the congruence measure was required for only 32% (56) of the descriptions. Correlations between judges’ assigned scores across descriptions was +.92 for the completeness
measure, +.71 for the accuracy measure, and +.76 for the congruence measure. These correlations are based on whole descriptions as the unit of analysis rather than features. Thus, it is possible for both judges to assign a description accuracy score of 3.5 to a description although disagreeing at the level of individual features. Such occurrences, however, were rare. More common was a criterion difference between judges in that one judge gave consistently lower accuracy and congruence scores than did the other judge across the entire sample of faces. Because of the generally high level of correlative agreement between judges across descriptions, however, it seems that the descriptions have been scored with reasonable reliability.

Effects of Pose, Target Presence, and Replication

An initial analysis of variance (ANOVA) was conducted to assess any differences between the four replication sets of target persons and the three-quarters versus straight-on pose. A $2 \times 4 \times 2$ (Straight-On Pose at acquisition and Three-Quarters Pose at identification vs. the Reverse Order × Replication × Target Present vs. Absent) on identification accuracy yielded a significant main effect for pose, $F(1, 160) = 3.42, p < .05$ (estimation $\rho^2 = 0.04$), and a significant main effect for target presence, $F(1, 160) = 35.57, p < .001$ (estimation $\omega^2 = 0.42$). Subjects were more likely to make a correct identification decision when they had the three-quarters pose at acquisition and the straight-on pose at identification than when they had the reverse order. Subjects also were more likely to make a correct identification decision when the target was present than when the target was absent. No other main or interaction effects were significant, $F$s $< 1.3$. The distribution of accurate responses and errors in target presence conditions was 15% identifications of the wrong photograph, 5% nonidentifications, and 81% accurate identifications. The distribution of accurate responses and errors in the target-absent conditions was 49% incorrect identifications and 51% correct nonidentifications.

Separate $2 \times 2 \times 4$ (Pose × Target Present vs. Absent × Replication) were conducted on the three description measures. Only one measure (number of features) yielded a significant effect which was a main effect for pose, $F(1, 160) = 7.64, p < .05$ (estimation $\omega^2 = .05$). The three-quarters pose at acquisition produced more description features than did the straight-on pose. This effect obviously could not be due to pose at identification, nor would there be any logical reason to expect target presence versus absence to affect the description measures as these factors were subsequent to obtaining descriptions.

Relationships Between Identification and Description

The main hypothesis concerns the relationship between description measures and identification accuracy. The correlations among these measures are reported in Table 1. Description completeness does not correlate significantly with either description accuracy or description congruence. This is not surprising as previous research indicates that completeness and accuracy of eyewitness reports are either unrelated or negatively related (e.g., Marquis, Marshall, & Oskamp, 1972). The fact that description accuracy and description congruence are correlated strongly is to be expected. Fifty-six percent of their data points are identical owing to the fact that congruence and accuracy are defined identically whenever the identified target is the correct target (which was true in 71 of the total 127 identifications).

More interesting is the fact that identification accuracy among the 127 identifiers was correlated significantly with description accuracy and also was correlated significantly with congruence of description.

Table 2 presents the data in another form that includes those who made no identification. These data show in another form that accurate identifiers gave more accurate descriptions on the average than did inaccurate identifiers. In addition, it appears that accu-

---

1 The effect for pose could be due to the relative efficacy of three-quarters pose versus straight-on pose at acquisition or the superiority of straight-on pose versus three-quarters pose at identification or some combination of these. As discussed previously, however, the purpose of the pose variable was not to assess its role in identifications but instead to make the task one of person identification rather than photograph identification.
Table 1
Correlations Between Description and Identification Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description completeness</th>
<th>Description accuracy</th>
<th>Description congruence</th>
<th>Identification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collapsed over conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description completeness</td>
<td>X</td>
<td>-.11</td>
<td>-.03</td>
<td>.05</td>
</tr>
<tr>
<td>Description accuracy</td>
<td>X</td>
<td>X</td>
<td>.87**</td>
<td>.27**</td>
</tr>
<tr>
<td>Description congruence</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>.19*</td>
</tr>
<tr>
<td>Identification accuracy</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

* This includes only subjects who made an identification (n = 127). Correlations in this column are point-biserial because accuracy is a dichotomous variable.
* p < .05. ** p < .01.

rate rejection subjects gave more complete and accurate descriptions than did inaccurate rejection subjects. However, the inaccurate rejections were too infrequent to subject them to analyses. Note also that description congruence cannot be calculated for any of the subjects who failed to make an identification (see earlier definition of congruence).

In order to help determine whether the relationship between descriptions and identifications is due to (a) subjects who are better at describing faces also are better at identifying faces or (b) faces that are easier to describe also are easier to identify, another analysis was conducted. Recall that each of the 88 target faces was described and identified by 2 subjects (one in each profile order). This analysis involved an intentional mixing wherein one subject's description of a given face was paired with the second subject's identification of that same face. The correlation between description accuracy and identification accuracy was then recalculated. The resulting correlation (r = .25) was nearly identical to that obtained when a subject's own description accuracy was used to predict the identification accuracy. Therefore, the description-accuracy correlation must be due to characteristics of the target faces.

Discussion

The results of this study give some support to the Supreme Court's (Neil v. Biggers, 1972) hypothesis that a prior description of a perpetrator has some utility for judging the accuracy of a subsequent identification. Technically, the hypothesis is consistent with the obtained data, but the low magnitude of the relationship raises questions about the practical utility of descriptions. Although there was .27 correlation between description accuracy and identification accuracy, the courts must deal instead with description congruence, which correlated only .19 with identification accuracy. A correlation of .19 accounts for less than 4% of the variance. Further caution may be required when considering the likelihood that the prediction variable (congruence) may be assessed less reliably by jurors than it was in this research which in turn would lower the relationship between congruence and identification accuracy.

These results raise questions about the process through which descriptions and identifications are related. Recall that the study by Wolfskiel and Brigham (in press) found no description-identification relationship even though they used a comparable number of subjects. It was suggested, however, that the description-identification relationship may result from a process wherein target faces that are easy to describe are also easy to identify. The Wolfskiel and Brigham study could not accommodate this hypothesis, however, but instead tested whether subjects

2 Wells and Lindsay (1985) point out, however, that the squared point-biserial correlation can mislead researchers into underestimating the practical utility of predictive relationships. This is especially true if the dichotomous criterion variable (identification accuracy in this case) has a base rate in the area of 50%. Furthermore, increases in interjudge reliability beyond the .76 obtained for the congruence measure might also increase the magnitude of the description congruence/identification accuracy correlation.
who are able to describe a face well are also able to identify that face. By using a wide sample of target faces, the current study allowed any systematic variance that was due to target faces to contribute to the description–identification relationship. In order to test whether the between-faces versus within-faces analysis was critical to the results, an additional analysis was conducted in which one subject's description of a given target face was used to predict another subject's identification of that same target face. The fact that this between-subjects analysis yielded correlations of similar magnitude to the within-subjects analysis indicates that it was the target faces themselves that accounted for the obtained relationship between description accuracy and identification accuracy.

Given the current results and the findings of Wolfskiel and Brigham (in press), it appears that the relationship between descriptions of faces and the identification of faces is not due to a process wherein good describers are good identifiers. Instead, it is due to a process wherein faces that are described easily are identified easily. Indeed, when trying to predict a given subject's identification accuracy for a given face, there is no advantage for using that subject's description score rather than some other subject's description score for that same face. It is likely that this is due to variance in faces along dimensions of uniqueness or typicality, which are thought to account for differences in face recognition (e.g., Courtois & Mueller, 1981; Goldstein & Chance, 1981; Light, Kayra-Stuart, & Hollander, 1979). Because the description–identification relationship is due to target factors rather than subject–witness factors, it seems reasonable to suggest that descriptions merely are poor correlates of some other variable that would better predict identification accuracy.

A major point of this research is that a relationship exists between descriptions of faces and identifications of faces and this relationship owes to target-face characteristics. However, the description–identification relationship was impressively small. Recent research suggests that this may be due to the way in which encoding and retrieval operations interact in memory for faces. Wells and Hryciw (1984) found that holistic encoding conditions (i.e., making personality trait judgments) facilitated identification of faces but yielded poor Identi-kit reconstructions; particularistic encoding conditions (i.e., judgments of features such as nose, eyes, etc.) yielded poor identification of faces but good Identi-kit reconstructions. Thus, a given encoding strategy may facilitate one form of retrieval (such as verbal retrieval) but not another (e.g., identification) which may lead to independence between some retrieval tasks.

3 The Identi-kit is a commercially available face-construction package consisting of transparencies of facial features (e.g., 34 separate noses, 106 pair of eyes) that can be superimposed over each other to construct a face.

References


Received February 24, 1984

Revision received January 10, 1985

New Look for the APA Journals in 1986

Beginning in 1986, the APA journals will have a new look. All the journals will be 8½ × 11 inches—a little larger than the *American Psychologist* is now. This change in trim size will help reduce the costs of producing the journals, both because more type can be printed on the larger page (reducing the number of pages and amount of paper needed) and because the larger size allows for more efficient printing by many of the presses in use today. In addition, the type size of the text will be slightly smaller for most of the journals, which will contribute to the most efficient use of each printed page.

These changes are part of continuing efforts to keep the costs of producing the APA journals down, to offset the escalating costs of paper and mailing, and to minimize as much as possible increases in the prices of subscriptions to the APA journals.