Running head: VERBAL REPORTS IN LINEUPS AND SHOWUPS

The Effects of Verbal Descriptions on Performance in Lineups and Showups

Brent M. Wilson[1]

University of California, San Diego

Travis M. Seale-Carlisle[1]

Royal Holloway, University of London

Laura Mickes

Royal Holloway, University of London

Word count: 7,263

Author Note

[1]Both authors contributed equally.

We thank John T. Wixted for discussions of this research.

Some of the data in this manuscript were presented at the Psychonomic Society 52nd Annual Convention.

Correspondence concerning this article should be addressed to Laura Mickes, Department of Psychology, Royal Holloway, University of London, Phone: +44 (0)1784 433711, Email: laura.mickes@rhul.ac.uk

This work was supported in part by the Economic and Social Research Council [ES/L012642/1] to Laura Mickes.

Abstract

Verbally describing a face has been found to impair subsequent recognition of that face from a photo lineup, a phenomenon known as the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990). Recently, a large direct replication study successfully reproduced that original finding (Alogna et al. 2014). However, in both the original study and the replication studies, memory was tested using only target-present lineups (i.e., lineups containing the previously-seen target face), making it possible to compute the correct ID rate (i.e., the hit rate) but not the false ID rate (i.e., the false alarm rate). Thus, the lower correct ID rate for the verbal condition could reflect either reduced discriminability or a conservative criterion shift relative to the control condition. In four verbal overshadowing experiments reported here, we measured both correct ID rates and false ID rates using photo lineups (Experiments 1 and 2) or single-photo showups (Experiments 3 and 4). The experimental manipulation (verbally describing the face or not) occurred either immediately after encoding (Experiments 1 and 3) or 20-minutes after encoding (Experiments 2 and 4). In the immediate condition, discriminability did not differ between groups, but in the delayed condition, discriminability was lower in the verbal description group (i.e., a verbal overshadowing effect was observed). A fifth experiment found that the effect of the immediate-vs.-delayed manipulation may be attributable to a change in the content of verbal descriptions, with the ratio of diagnostic to generic facial features in the descriptions decreasing as delay increases.

Keywords: verbal overshadowing effect; lineups; showups; discriminability; reliability

Introduction

A police lineup is administered to victims and eyewitnesses to aid criminal investigations. The lineup is a collection of individuals, including the police suspect (who may be innocent or guilty) and a number of fillers (who are known to be innocent and resemble the perpetrator). Verbally reporting the details of a crime is a necessity in the investigative process. Whether or not the very act of reporting details about the perpetrator retrieved from memory impairs later memory for the perpetrator has been a topic of interest and debate for the last several decades. Interest in this topic was triggered by a finding reported by Schooler and Engstler-Schooler (1990) in which participants watched a video of a simulated robbery and either verbally described the perpetrator or engaged in a control task. Participants who gave verbal descriptions were significantly less likely to correctly identify the perpetrator from a lineup test than those in the control condition. This somewhat counterintuitive finding, termed the "verbal overshadowing effect," has potential implications for the criminal justice system.

Because follow-up research yielded mixed results and a meta-analysis yielded effect sizes much smaller than the original experiments (Meissner & Brigham, 2001), a large direct replication study was recently conducted on two of the original experiments (Experiments 1 and 4 of Schooler & Engstler-Schooler, 1990; Alogna et al. 2014). In both experiments, the main experimental manipulation was the same: participants either verbally described the perpetrator or took part in a control task. The only difference between the experiments was the order of procedural events. As shown in Figure 1, the experimental manipulation took place immediately after presentation of the video in Experiment 4 of Schooler and Engstler-Schooler (Figure 1A) or 20 minutes after the presentation of the video in Experiment 1 of Schooler and Engstler-Schooler (Figure 1B). The effect replicated: compared to the control condition, the perpetrator was less likely to be identified from

the lineup in both experiments, but the effect was much larger when the verbal description was provided 20 minutes after the video (and immediately before the lineup test).

In a typical eyewitness identification study, some participants are presented with a target-present lineup (i.e., a lineup that contains a photo of the guilty suspect) and other participants are presented with a target-absent lineup (i.e., a lineup in which the photo of the guilty suspect has been replaced by a photo of the innocent suspect). The measures of interest are the correct ID rate (the proportion of participants presented with a target-present lineup who correctly identify the guilty suspect) and the false ID rate (the proportion of participants presented with a target-absent lineup who incorrectly identify the innocent suspect). However, the original verbal-overshadowing experiments and the studies that recently replicated them included target-present lineups (i.e., lineups that contained a photo of the guilty suspect), but did not include target-absent lineups. What these studies therefore showed is that the correct ID rate was lower in the verbal description condition compared to the control condition. The effect of that manipulation on the false ID rate is unknown. Thus, the only safe conclusion is that there was a reduction in the probability of correctly identifying the perpetrator when a verbal description was provided, but whether that reduction in the correct ID rate occurred because of reduced discriminability or because of a more conservative response bias is unknown. Distinguishing between those alternative interpretations requires that the probability of identifying the innocent suspect be measured as well (e.g., Clare & Lewandowsky, 2004, Mickes, 2016; Mickes & Wixted, 2015; Rotello, Heit, & Dube, 2015; Smith & Flowe, 2015). Moreover, the applied implications of the verbal overshadowing effect are fully dependent on whether it arises because of reduced discriminability or because of a conservative response bias (Mickes & Wixted, 2015).

Does providing a verbal description reduce discriminability?

The "verbal overshadowing effect" refers to impaired recognition memory performance. Interpreted in terms of signal detection theory, impaired recognition performance refers to reduced discriminability. Thus, from that perspective, a true verbal overshadowing effect is properly defined as a reduction in discriminability – that is, a reduction in the ability to discriminate innocent from guilty suspects – as a consequence of describing the perpetrator (Mickes, 2016). To measure discriminability, both the correct ID rate and false ID rate must be taken into account.

To measure discriminability, the most accurate approach is to measure not just one correct and false ID rate per condition (e.g., verbal description vs. control) but to measure the full range of correct and false ID rates that can be achieved in each condition across different levels of response bias. The entire family of achievable correct and false ID rates for a given condition is known as the receiver operating characteristic (ROC). ROC analysis is most easily performed by plotting correct vs. false identification rates across different levels of confidence. The ensuing ROC curves are constructed for both conditions and the area under the curve (AUC) for each condition is measured and statistically compared (for descriptions of how to conduct ROC analysis of lineup data, see Gronlund, Wixted & Mickes, 2014; Mickes, Flowe & Wixted, 2012). The larger the AUC, the better the discriminability. Evidence of a true verbal overshadowing effect would consist of a smaller AUC when a verbal description is provided compared to when a verbal description was not provided.

Does providing a verbal description affect reliability?

If verbal overshadowing does in fact reduce discriminability, it seems natural to suppose that it reduces the reliability of a suspect ID. However, whether or not providing a verbal description affects discriminability is a different question than whether or not providing a verbal description affects the *reliability* of a suspect identification from a lineup. Whether discriminability is low or high, an experimental manipulation that induces conservative responding will yield relatively high reliability (i.e., identifications will tend to be accurate), whereas a manipulation that

induces liberal responding will yield relatively low reliability (i.e., identifications will tend to be less accurate). Reliability can be measured in several different ways, including calibration analysis and confidence-accuracy characteristic (CAC) analysis (Mickes, 2015). Calibration analysis has the potential to underestimate reliability because the relevant equation includes filler identifications (for a comprehensive explanation of the differences between calibration and CAC analysis, see Wixted, Read, & Lindsay, 2016). CAC analysis, on the other hand, involves only suspect (guilty and innocent) ID accuracy as a function of confidence. This is the measure that is of most relevance to the legal system, which is interested in knowing the probability that a suspect who has been identified is actually guilty. When the base rates of target-present and target-absent lineups are equal (as is typically true of lab studies), this measure is given by:

$$PPV = \frac{S_g}{S_g + S_i}$$

where PPV is positive predictive value, S_g is the number of correct identifications, and S_i refers to the number of estimated innocent suspect identifications[1]. PPV is the probability that a suspect who was identified by a witness is in fact guilty, and it is computed separately for every level of confidence. For example, for participants who identify a suspect with high confidence, the PPV_{high} is given by:

$$PPV_{(high)} = \frac{S_{g(high)}}{S_{g(high)} + S_{i(high)}}$$

where $S_{g(high)}$ is the number of correct suspect IDs made with high confidence and $S_{i(high)}$ is the number of innocent suspect identifications made with high confidence. If the PPV for identifications made with high confidence were higher when no verbal description was provided, then reliability would be higher in that condition. It is possible, however, that even if discriminability is lower when

[1] If there is no designated innocent suspect, the false suspect identification rate is estimated by dividing filler IDs from target-absent lineups by the number of lineup members.

a verbal description is provided, reliability could be higher (Mickes, 2016). This could happen if, for example, verbal descriptions reduced discriminability while at the same time induced very conservative responding.

Investigating the Effect of Verbal Descriptions on Discriminability and Reliability

To test the effect of verbal descriptions on discriminability and reliability, we directly replicated Schooler and Engstler-Schooler (1990) in four experiments. Following suit of the replication studies (Alogna et al., 2014), we replicated Experiments 1 and 4 of the original paper, with one critical difference – the inclusion of target-absent lineups. The original and replication studies tested memory on 8-person simultaneous lineups. To be able to use the same stimuli and include target-absent lineups, the lineup size was reduced to 6-person simultaneous lineups so that the perpetrator could be replaced with a filler for target-absent lineups. In our Experiment 1, the experimental manipulation (verbal description vs. control task) took place immediately after the study phase, and in our Experiment 2, the experimental manipulation took place 20 minutes after the study phase (Figure 1).

In two additional experiments, we tested the effect of verbal descriptions on showups. Showups involve the presentation of only one person (the suspect) on the recognition test. Though showups are believed to be highly suggestive in nature (Goodsell, Wetmore, Neuschatz, & Gronlund, 2013; Steblay, Dysart, Fulero, & Lindsay, 2003) and have been found to yield lower discriminability than lineups (Wetmore et al., 2015; Mickes, 2015), showups will continue to be widely used by the police because they can be administered soon after a crime has been committed. As in Experiments 1 and 2, Experiments 3 and 4 retained the same procedural order as the original and replication experiments (see Figure 1). For these experiments, discriminability and reliability were again measured with ROC and CAC analysis, respectively. Finally, in a fifth

experiment, we conducted a content analysis in an effort to determine why verbal descriptions have the effect they do on recognition memory performance.

Experiment 1

Method

Participants

Undergraduate students (*N* = 780) at the University of California, San Diego (UCSD) participated online for course credit. Sample size (for Experiments 1 and 2) was based on a power calculation that aimed to achieve 80% power (using results from an earlier lineup study, Mickes, Flowe, & Wixted, 2012, to estimate the effect size). Participants (*n* = 63) reported that they previously viewed the video and were therefore not included in the analyses. Of the remaining (*n* = 717; 472 female, 239 male, and 6 did not specify), the average age = 20.5 years (*sd* = 2.55). Participants were randomly assigned to the control condition or the verbal condition and were tested on a target-absent lineup ($n_{control}$ = 188; n_{verbal} = 168) or a target-present lineup ($n_{control}$ = 171; n_{verbal} = 190) based on random assignment. The UCSD Institutional Review Board approved all of the experiments.

Materials

The stimuli included the 44-second video of the mock bank robbery and the eight photos (one of the perpetrator and seven fillers) used in original experiments (Schooler & Engstler-Schooler, 1990). The test phase included 6-person lineups with the images arranged in a 2x3 array. Images of the target and fillers from the original experiments were used for target-present and target-absent lineups. Target-present lineups were constructed using five of the seven fillers' images (that were randomly selected for each participant), and the photo of the perpetrator (and all of the images were randomly arranged for each participant). Target-absent lineups were

constructed using six of the seven fillers' images (that were randomly selected and randomly arranged for each participant). The distractor task was an online crossword puzzle similar to the puzzle used in the original experiments (Schooler & Engstler-Schooler, 1990).

Procedure

The experiment was conducted online. Participants watched the video, typed as many countries and capitals as possible within 5-minutes (control condition) or provided a description of the perpetrator from the video for 5-minutes (verbal condition), and engaged in the 20-minute distractor task (see Figure 1A). The same instructions listed in the approved final protocol for the Alogna et al. (2014) study were used for the 5-minute writing task for participants in both conditions. Those in verbal conditions were given the following instructions from Alogna et al.: "Please describe the appearance of the bank robber in as much detail as possible. It is important that you attempt to describe all of his different facial features. Please write down everything that you can think of regarding the bank robber's appearance. It is important that you try to describe him for the full 5 minutes" (pp. 559-560). After a 20-minute distractor task, memory for the perpetrator was tested on a lineup where participants were asked to try to identify the perpetrator or choose the "not present" option and rate their confidence on a 7-point scale (1 = guessing; 7 = certain).

Results and Discussion

Table 1 shows the frequency counts for each response type for target-present and target-absent lineups by levels of confidence. The average correct ID rate was higher for the verbal condition (0.52) than the control condition (0.49). False ID rates were estimated by dividing the number of filler identifications by the number of lineup members. The estimated false ID rate was lower for the verbal condition (0.09) than the control condition (0.11). Thus, discriminability was, if

anything, higher in the verbal group. We conducted ROC analysis to compare the locus of correct and false ID rates of both groups for a more complete assessment. The ROC curves in Figure 2A show higher discriminability for the verbal group. However, using a false ID rate cutoff of .458,[2] partial area under the curve (*pAUC*) analyses revealed that the difference between the verbal group (0.183) and the control group (0.153) was not significant, $D = 1.20$, $p = .232$. In other words, neither a verbal overshadowing effect nor its opposite was observed.

The curves in Figure 2A were generated from fitting a basic equal variance signal detection-based model to the ROC data from the control and the verbal conditions. In the model, memory strengths are distributed according to two Gaussian distributions, one representing fillers (which includes an innocent suspect) and one representing targets. This model assumes that to create target-absent lineups, six random draws are made from the filler distribution, and to create target-present lineups, five random draws are made from the filler distribution and one random draw is made from the target distribution. In the simplest version of this model, an identification is made if the memory strength of the most familiar face in the lineup exceeds a decision criterion.

The filler distribution was set to $\mu_{lure} = 0$, $\sigma_{lure} = 1$, and the corresponding mean for the target distribution was estimated by fitting the model to the data. Correct and false identifications were binned into low (ratings of 1-3), medium (ratings of 4-5) and high (ratings of 6-7) levels of confidence and used for the different decision criteria. The model estimates d' and the three decision criteria. Fits were improved (but conclusions were not changed) by including another parameter, δ, which scales the estimated placements of the confidence criteria for target-present lineups relative to target-absent lineups (Seale-Carlisle & Mickes, 2016). Thus, there were a total of

[2] This value was selected because it is the false ID rate of the rightmost point on the ROC curve of the verbal condition, the more conservative of the two conditions (Gronlund, Wixted, & Mickes, 2014). Using the false ID value of the rightmost point on the ROC curve of the control condition as the cutoff does not change the conclusion ($p = .142$).

10 parameters for both conditions, and each condition had 18 degrees of freedom: 3 degrees of freedom (filler identifications made with low, medium, or high confidence) for target-absent lineups (both conditions) and 6 degrees of freedom (filler identifications or suspect identifications made with low, medium, or high confidence) for target-present lineups (both conditions). The fits were performed simultaneously and had 8 degrees of freedom (18 degrees of freedom - 10 free parameters).

The parameters were adjusted until the difference between observed and predicted frequency counts was minimized using a chi-square goodness-of-fit statistic. The fit was good, $\chi^2(8) = 8.41$, $p = .394$. Constraining d' to be equal for verbal and control conditions also resulted in a good fit, $\chi^2(6) = 10.13$, $p = .119$. The full model fit and the constrained model fit did not differ significantly, $p = .190$, indicating that d' did not differ for the two conditions. Thus, as is typically (but not necessarily) true, the results from the atheoretical *pAUC* analysis and the theoretical signal detection analysis agree.

The analyses presented above were concerned with discriminability. As noted earlier, reliability is a different issue. To measure the reliability of suspect IDs as a function of confidence in each condition, we conducted CAC analysis. Confidence ratings were binned in the same manner as for the model fits, and CAC was computed for identifications made with low, medium and high levels of confidence. The error bars represent standard error bars estimated using a bootstrap procedure (see Seale-Carlisle & Mickes, 2016). Figure 2B shows that the verbal group had higher reliability at each level of confidence, but none of the differences were significant.

Overall, neither discriminability nor reliability differed significantly between groups. In the original experiment in which the experimental manipulation occurred immediately after the study phase and 20 minutes prior to the identification procedure, the correct ID rate for the control group (0.71) was much higher than the verbal group (0.49) (Schooler & Engstler-Schooler, 1990). In

the analogous replication experiment (Alogna et al., 2014), the average correct ID rate for the control group (0.55) was also higher than the verbal group (0.51), but the difference between the correct ID rates reported was considerably smaller. Three of the 31 participating laboratories found no difference between conditions, and 10 found a higher correct ID rate in the verbal condition than in the control condition. Our results also revealed slightly higher correct ID rates for the verbal group (0.52) than the control group (0.49).

Confidence and accuracy were related for both groups. Identifications made with medium confidence were higher in accuracy than identifications made with low confidence, and lower in accuracy than identifications made with high confidence. Furthermore, CAC analysis revealed that identifications made with high confidence were comparably reliable for both groups.

Experiment 2

Experiment 2 was the same as Experiment 1 with the exception of swapping procedural order. In Experiment 2, the experimental manipulation took place 20 minutes after the study phase and immediately before the identification test (see Figure 1B). This was the order in which the greatest difference in correct identification rates resulted between groups in the replication experiments (Alogna et al., 2014). Also, as in Experiment 1, target-absent lineups were included to assess discriminability and reliability.

Method

Participants

Participants (*N* = 780) were recruited from Royal Holloway, University of London (*n* = 138), Amazon Mechanical Turk (*n* = 245), and SampleSize (*n* = 397). The participants (*n* = 10) who reported previously viewing the video were excluded from the analyses. Of the remaining (*n* = 770,

442 female; 318 male; 10 did not state), the average age = 27.9 years (*sd* = 11.1) Participants were randomly assigned to the control condition or the verbal condition and a target-absent lineup ($n_{control}$ = 179; n_{verbal} = 185) or a target-present lineup ($n_{control}$ = 196; n_{verbal} = 210). Royal Holloway, University of London Research Ethics Committee approved this study.

Materials and Procedure

The materials were the same used in Experiment 1. The procedure was the same with one exception: the experimental manipulation took place after the 20-minute distractor task and immediately before the test phase (see Figure 1B).

Results and Discussion

Table 1 shows the frequency counts for target-present and target-absent lineups by levels of confidence. The correct ID rate was lower in the verbal group (0.38) compared to the control group (0.62). The false ID rate was also lower in the verbal group (0.07) than the control group (0.09), which could mean that there is a difference in response bias, not discriminability, per se. We therefore conducted ROC analysis to measure discriminability independent of response bias. Figure 3A shows the ROC curves for both groups, and discriminability was lower in the verbal group than the control group. Using a false ID rate cutoff of .584,[3] *pAUC* analysis revealed that the difference between the verbal (.096) and control (.155) groups was significant, D = 3.06, p = .002.

The ROC curves were generated from the same equal variance signal detection model as in Experiment 1. The ROC data were also fit, using the same parameters as in Experiment 1, and again, the fit was good, $\chi^2(8) = 6.12$, p = .634. However, when d' was constrained to be equal, the fit was worse, $\chi^2(6) = 17.53$, p = .008, and the fit was significantly different than when d' values were free

[3] Consistent with Experiment 1, this value was selected because it is the rightmost point on the ROC curve of the verbal condition (the more conservative condition). Using the false ID value of the rightmost point on the ROC curve of the control condition as the cutoff does not change the conclusion (p = .002).

to vary, $p < .001$. Thus, once again (and as expected), atheoretical *pAUC* analysis and theoretical signal detection analysis agree that discriminability was reduced in the verbal condition.

We next turned to the issue of reliability. The CAC curves, shown in Figure 3B, show slightly lower reliability across all three levels of confidence for the verbal group, but the differences were not significant at any of the levels of confidence. Moreover, confidence and accuracy are related (i.e., high confidence identifications are more accurate than low confidence identifications). High-confidence accuracy in the control condition was .95, whereas high-confidence accuracy in the verbal condition was .91. Thus, in both conditions, accuracy was high, and the small difference between them was not significant.

In the current experiment, the correct ID rate was lower in the verbal condition compared to the control condition (0.38 vs. 0.62, respectively). This pattern was consistent with the original experiments (Schooler & Engstler-Schooler, 1990) and replication experiments (Alogna et al., 2014) when the verbal description task was delayed for 20 minutes after encoding. In the former, the correct ID rate was lower for the verbal condition (0.39) vs. the control condition (0.64). Similarly, in the latter, the average correct ID rate for the verbal condition (0.38) was lower than that of the control condition (0.54). Although those results are ambiguous as to whether they reflect either reduced discriminability or more conservative responding (or both), the ROC results reported here revealed significantly lower discriminability in the verbal condition (Figure 3A).

Despite the fact that discriminability was lower in the verbal condition, reliability was not significantly different (similar to the findings in Experiment 1) between the two conditions. Furthermore, high-confidence identifications were much more accurate than low-confidence identifications in both conditions. Thus, with regard to assessing the probative value of an ID, knowing confidence is far more informative than knowing whether or not the suspect's face was verbally described (despite the large verbal overshadowing effect). Next, we extended the

replication further by testing the effect of verbal descriptions on discriminability and reliability when memory is tested using showups.

Experiment 3

In Experiment 3 we sought to replicate the pattern of results from Experiment 1. Thus, the procedure was held constant except that participants were tested on either a target-present or target-absent showup (i.e., the guilty suspect or innocent suspect, respectively). Again, we measured discriminability with ROC analysis and reliability with CAC analysis.

Method

Participants

UCSD undergraduate students participated online for course credit (N = 1,197; 410 male, 773 female, 14 unspecified; average age = 20.2 years, sd = 2.7). There are no earlier showup studies (i.e., showup vs. showup studies) to inform a power analysis, so sample size (for Experiments 3 and 4) was increased to 1,100 and we stopped data collection when the term ended. Participants were randomly assigned to the control condition or the verbal condition. Participants were also randomly assigned to a target-absent showup (control n = 300; verbal n = 293) or a target-present showup (control n = 328; verbal n = 276).

Materials

The materials were the same as those in Experiment 1 and 2, except an online game of Tetris was played instead of a crossword puzzle as the distractor task. Target-present showups were constructed by using the target photo, and target-absent showups were constructed by randomly selecting one of the seven filler photos.

Procedure

Procedural order was the same as in Experiment 1 (see Figure 1A). The only differences were that showups replaced lineups and participants rated their confidence on a 0-100% scale (0 = guessing; 100% = certain).

Results and Discussion

Table 1 shows the frequency counts for each response type for target-present and target-absent lineups by levels of confidence. The correct ID rate was higher for the control condition (0.65) than for the verbal condition (0.57). Likewise, the false ID rate was higher for the control condition (0.29) than for the verbal condition (0.18). Thus, on the surface, these results indicate that, at a minimum, verbal descriptions induced more conservative responding. Figure 4A shows that the two conditions yielded ROC curves that are not noticeably different, suggesting that verbal descriptions did not affect discriminability. The statistical comparison of the *AUC* values between the verbal (0.756) and control (0.735) conditions confirm this impression, $D = 0.71$, $p = .481$.

Next, correct and false identifications were binned into low (0-60%), medium (70-80%), and high (90-100%) confidence ratings to perform signal detection model fits to the ROC data and to conduct CAC analysis. The ROC curves in Figure 4A were generated by fitting the ROC data using the same equal variance signal detection model described previously with one less parameter (because there are no filler identifications with showups). The fit was good, $\chi^2(6) = 6.76$, $p = .344$. Constraining d' to be equal did not significantly worsen the fit, $p = .663$.

The CAC curves (using the same confidence binning as for the model fits) in Figure 4B show no significant reliability differences between condition across the levels of confidence. Once again, confidence is predictive of accuracy, but the relationship for the verbal condition does not continue to increase from medium to high confidence. Also, high confidence identifications are noticeably lower in accuracy compared with what we observed for lineups (averaged across conditions, high-

confidence showup accuracy = 0.80).

Experiment 4

In Experiment 4 we sought to replicate the pattern of results from Experiment 2, and like in Experiment 3, memory was tested on a target-absent or target-present showup. Again, we measured discriminability with ROC analysis and reliability with CAC analysis.

Method

Participants

UCSD undergraduate students participated online for course credit (N = 1,196; 364 male, 822 female, 10 unspecified; average age = 20.3 years, sd = 2.3). Participants were randomly assigned to the control condition or the verbal condition. Memory was tested on a target-absent showup ($n_{control}$ = 302; n_{verbal} = 322) or a target-present showup ($n_{control}$ = 311; n_{verbal} = 261).

Materials

All materials were the same as in Experiment 3.

Procedure

The procedure was the same as Experiment 3 with exception of the order of the distractor task and the experimental manipulation (the same order as Experiment 2; see Figure 1B). The writing task took place after the 20-minute distractor task and immediately before the test phase.

Results and Discussion

Table 1 shows the frequency counts for each response type for target-present and target-absent lineups by levels of confidence. Similar to the results in Experiment 2 (and the analogous replication study), the correct ID rate was lower in the verbal condition (0.43) than in the control condition (0.68). Also as in Experiment 2, the false ID rate was lower in the verbal condition (0.17)

than the control condition (0.25). Again, these results are consistent with the idea that, at a minimum, providing verbal descriptions induced more conservative responding. To measure discriminability, ROC analysis was conducted. The ROC curves, as shown in Figure 5A, and AUC analysis reveal that discriminability is lower for the verbal condition (0.70) than the control condition (0.77), and that difference is significant, $D = 2.36$, p = .018. The curves in Figure 5A were generated by fitting the ROC data using the same equal variance signal detection model described in Experiment 3. Again, the fit was good, $\chi^2(6) = 3.28$, $p = .778$, and constraining d' to be equal worsened the fit to a marginally significant degree, $p = .052$. Thus, a verbal overshadowing effect is evident whether an atheoretical measure (AUC) or a theoretical measure (d') is used to interpret the results.

Despite the difference in discriminability, the CAC curves in Figure 5B again reveal no significant differences between conditions in reliability at all levels of confidence. However, there is a trend towards lower accuracy in the verbal conditions for IDs made with low or medium confidence. As in the other experiments, identifications made with high confidence are higher in accuracy than identifications made with medium and low confidence. High-confidence accuracy was 0.87 in both conditions. The results shown in Figure 5A and 5B illustrate a key point: a reduction in discriminability does not automatically translate into reduced reliability of IDs made with high confidence. As noted earlier, knowing the effect of a variable on discriminability does not automatically reveal the effect of that same variable on the reliability of an ID.

Experiment 5

In Experiments 1 and 3, participants in the verbal condition provided descriptions immediately after encoding, and discriminability did not differ from the control condition, regardless of whether memory was tested using a lineup or a showup. However, when verbal

descriptions were provided after a delay (as in Experiments 2 and 4), discriminability was impaired for both procedures. What accounts for the difference in discriminability depending on whether or not the description is delayed?

The diagnostic feature-detection hypothesis may provide insight into this difference (Wixted & Mickes, 2014). The hypothesis was initially proposed to account for the discriminability advantage that simultaneous lineup presentations have over procedures that involve showing an individual in isolation (as with sequential lineups or showups). By seeing the lineup members together, it is readily apparent to the witness that there are facial features shared across lineup members that should be discounted because they are not diagnostic of guilt. For example, if the perpetrator were a young, White male, then attaching weight to those features would not be helpful and would instead serve to impair discriminability because all of the lineup members would be young, White males. Having the faces presented simultaneously allows eyewitnesses to immediately detect and discount non-diagnostic features and to instead attach more weight on features that are not shared and are thus more diagnostic. This discrimination-enhancing strategy is less likely to be used when lineup members are presented individually because, under those conditions, it is harder to detect (and then discount) the common, non-diagnostic facial features.

The same concept may help to explain why verbal descriptions only impair discriminability when they are made after a delay. More specifically, participants may use more diagnostic feature descriptions immediately after encoding the perpetrator's face than they do after a delay. After a delay, by contrast, some forgetting will undoubtedly occur, and the description may become more general, perhaps becoming more likely to correspond to the common features that match everyone in the subsequently presented lineup. In that case, the participants may have a tendency to rely on the description they just gave when trying to identify the face of the perpetrator. To the extent that

they rely on the general (common) facial features mentioned in the verbal description, discriminability would be impaired.

To assess whether or not the diagnostic feature-detection hypothesis can help to account for the differences in discriminability when verbal descriptions are delayed, we first conducted a content analysis of the verbal descriptions provided in Experiments 1 through 4. We then conducted an experiment to test our theory.

Content Analysis

To conduct content analysis, 20 words were identified based on the appearance of the eight images of the perpetrator and fillers. Ten words were selected that were judged by the experimenters to be useful in differentiating the perpetrator from fillers (diagnostic-feature words), and 10 words were selected that were also judged by the experimenters to be less useful in differentiating the perpetrator from fillers (non-diagnostic-feature words). The latter words could have been used when selecting the fillers (e.g., White, male, attributes that related to hair color and stature). The diagnostic-feature words were descriptors that were not shared by all of the lineup members (see Appendix). The non-diagnostic-feature words were descriptors that were shared by all of the lineup members (see Appendix). The diagnostic-feature and non-diagnostic feature words were counted from descriptions provided by participants in the verbal condition in Experiments 1 and 3 (immediate descriptions) and compared with those descriptions in Experiments 2 and 4 (delayed descriptions).

Significantly more diagnostic-feature words were used when verbal descriptions were provided immediately after encoding (Experiments 1 and 3) compared to when verbal descriptions were provided 20 minutes after encoding (Experiments 2 and 4), $t(1945) = 4.75$, $p < .001$, Cohen's $d = 0.22$. However, there was no significant difference between the number of non-diagnostic feature words, $t(1945) = 1.28$, $p = .201$. A 2 × 2 analysis of variance revealed a significant interaction

between type of feature (diagnostic vs. non-diagnostic) and time of verbal description (immediate vs. delayed), $F(1, 3890) = 15.96$, $p < .001$, Cohen's $d = 0.06$. These results provide evidence for the diagnostic feature-detection hypothesis.

In light of these findings, we conducted an experiment to test whether more diagnostic words were used when the verbal descriptions were provided immediately after encoding compared to after a delay. The participants in this experiment did not watch a video of the perpetrator but instead read either the descriptions that were written immediately after encoding or after a delay. They were then tasked with trying to identify the perpetrator from a lineup based on description only (i.e., they did not view the video). If the immediate descriptions contain more diagnostically useful information, then participants provided with those descriptions should be better able to identify the perpetrator from the lineup compared to the participants provided with the delayed descriptions.

Method

Participants. UCSD undergraduate participants took part in exchange for course credit (N = 128; 44 male, 81 female, 3 unspecified; mean age = 20.3 years, sd = 2.3). There are no previous studies to inform a power analysis, so we selected a sample size of 100 and stopped collecting data at the end of the term. Participants were randomly selected to read descriptions from Experiment 3 (n = 63) or Experiment 4 (n = 65). None had participated in the previous experiments.

Materials. The materials were the descriptions written by the participants in the verbal condition in Experiment 3 (written immediately after encoding; Figure 1A) and Experiment 4 (written after a 20-minute delay; Figure 1B), which were 569 and 583 descriptions, respectively.

Procedure. For each participant, one description was randomly selected from the pool of descriptions. Participants read the description and were immediately presented with an 8-person

simultaneous target-present lineup (using the seven fillers and the perpetrator described in the previous experiments). The images were arranged in random order for each participant. Based on the description they read, participants attempted to identify the person they thought had committed the crime with no option to reject the lineup (i.e., no "not present" option).

Results and Discussion

Participants who read the descriptions that were written immediately after encoding were significantly more likely to correctly identify the perpetrator (M = 0.14, 95% CI = [.08, .25]) than participants who had read descriptions written after the delay (M = 0.03, 95% CI = [.01, .11]), z = 2.26, p = .024. Note that selecting a lineup member randomly from a perfectly fair 8-person lineup would result in a correct ID rate of 0.13. However, no lineup is perfectly fair. The low perpetrator selection rate in the control condition could either mean that the lineup was inherently biased towards one or more of the fillers (away from the perpetrator) or that the descriptions written after a delay had the effect of biasing selections towards one or more of the fillers (perhaps because they matched a more generic description than the perpetrator did).

In agreement with the diagnostic feature-detection hypothesis, more diagnostic-feature words were used in the descriptions when those descriptions were provided straightaway. Also consistent with the diagnostic feature-detection hypothesis, participants were able to identify the perpetrator more often if they read the description that was written by participants who provided the description immediately after encoding versus after a delay. Those descriptions therefore must have been more informative (i.e., more diagnostic). If participants in a verbal overshadowing experiment rely to some extent on their own descriptions when attempting to identify the perpetrator from the lineup, the prediction would be that discriminability should be impaired when descriptions are delayed (an effect that was observed in Experiments 2 and 4).

General Discussion

In a series of experiments we investigated the effects of verbal descriptions on discriminability and reliability on lineups and showups. The correct identification findings replicated the original verbal overshadowing Experiments 1 and 4 of Schooler and Engstler-Schooler (1990) and the replication efforts (Alogna et al., 2014). However, conclusions about memory performance based only on correct ID rates are tenuous. We therefore extended those findings by including target-absent lineups to be able to assess discriminability and reliability.

Effects of Verbal Reports on Discriminability

In Experiments 1-4, responding was more conservative in the verbal description condition. The relative conservatism is seen in the ROC curves in Figures 2-5 where the rightmost point on the verbal ROC is shifted leftward relative the rightmost point on the control ROC. Thus, one effect of providing verbal descriptions is to induce more conservative responding, and this phenomenon could account for the lower correct ID rates found in the original (Schooler & Engstler-Schooler, 1990) and the replication studies (Alogna et al., 2014). Why might this be? Clare and Lewandowsky (2004) proposed the idea that the task of describing the perpetrator makes participants realize that the task is challenging and as a result induces more cautious responding when faced with making a lineup decision[4]. Our findings are consistent with this idea. However, above and beyond the conservative shift in responding, the results of Experiments 2 and 4 (involving delayed verbal descriptions) showed that discriminability in the verbal condition was also impaired.

Why is discriminability impaired by providing a verbal description after a delay but not by providing a verbal description immediately? This puzzling difference in discriminability could be explained by the diagnostic feature-detection hypothesis (Wixted & Mickes, 2014). The hypothesis

[4] While Clare and Lewandowsky found lower discriminability (as measured by *d'*) for participants in one of their verbal conditions (the Holistic condition) compared to a control condition in Experiment 1 (the experiment most analogous to Experiments 1 and 2 here), they focused on the differences in criterion shifts.

holds that discriminability will be better when eyewitnesses rely more on diagnostic features than less diagnostic features. We tested this account in two ways, by conducting a content analysis and an experiment. In both analyses, the diagnostic feature-detection hypothesis provided a coherent interpretation of the data. Participants provided less diagnostic descriptions after a delay (presumably due to forgetting of more specific diagnostic details) compared to when descriptions were made immediately after encoding the face of the perpetrator, and other participants provided with those descriptions (but who did not see the mock-crime video) were better able to identify the perpetrator using the more diagnostic descriptions that had been written immediately after encoding.

Why people use less diagnostic information after time passes is a question that remains to be answered. Two potential theories may provide insight: fuzzy-trace theory and dual-process theories of recognition memory. Fuzzy-trace theory (Brainerd & Reyna, 1990) predicts that descriptions given after a delay would be based on gist representations versus descriptions given immediately, which would be based more on verbatim representations. This shift occurs because verbatim representations are thought to fade more rapidly than gist-based representations (e.g., Reyna, 2012). Indeed, Schooler (1998) once broadly linked verbal overshadowing with fuzzy trace theory, and it may be time to revisit this connection with more focus on the differential time course of gist and verbatim traces. Similarly, dual process theories might predict that descriptions provided soon after encoding are based on recollection, whereas descriptions provided later are based more on familiarity (e.g., Wais, Wixted, Hopkins, & Squire, 2006, but see e.g., Fortin, Wright, & Eichenbaum, 2004). Determining the usefulness of these theories could be a target for future research efforts.

One possibility that cannot be ruled out by our findings is that participants who provided a description immediately after encoding may rely on their description less than participants who

provided a description after a delay. Adding an additional 20-minute delay would be one way to assess this possibility. Another possibility that cannot be ruled out by our findings is that when the description is provided after a delay, participants rely less on diagnostic information despite the fact that the memory is intact and more diagnostic information can be culled in ways that were not tested here. Relatedly, future research efforts could involve investigations of ways to induce eyewitnesses to generate diagnostic descriptions even after time passes.

Effects of Verbal Reports on Reliability

In Experiments 1 through 4, the reliability of suspect IDs was comparable between conditions. Thus, even when discriminability was lower in the verbal condition, as was the case in Experiments 2 and 4, reliability was not appreciably different. Furthermore, adding to the body of literature that confidence and accuracy are related (e.g., Juslin, Olsson, & Winman, 1996; Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer, Brewer, Zweck, & Weber, 2010; Dodson & Dobolyi, 2016; Mickes, 2015; Wixted, Read, & Lindsay, 2016), the relationship was strong for both of the conditions in these experiments. That is, PPV for high confidence identifications was higher than PPV for medium confidence identifications, which was higher for low confidence identifications. This was true even when the effect of verbal overshadowing on discriminability was strong (Experiments 2 and 4).

A similar pattern (reduced discriminability without a concomitant reduction in reliability) has now been reported for manipulations such as retention interval (Palmer et al., 2013; Sauer et al., 2010; Wixted et al., 2016), same-vs.-cross race (Dodson & Dobolyi, 2016; Nguyen, Pezdek & Wixted, in press), and both exposure duration and divided attention (Palmer et al., 2013). In each case, the manipulation in question had a strong effect on discriminability while having little to no effect on the reliability of an ID made with high confidence. Although fewer high-confidence IDs

occur in the low-discriminability condition, when they do occur in that condition, they are typically as accurate (or nearly so) as high-confidence IDs in the high-discriminability condition.

Practical Implications

The implications of these results for the criminal justice system seem straightforward. The results from ROC and CAC analyses are of interest to different decision-makers with ROC analysis being important for policymakers, who decide whether and when to ask for a verbal description, and CAC analysis being important for judges and jurors, who have no control over police policy but ought to know how reliable an ID is likely to be (Mickes, 2015; Mickes, 2016).

Our ROC results suggest that, as they presumably already do, police should encourage reporting crimes immediately and then take down the description of the perpetrator as soon as possible. By doing so, the adverse effects of verbal descriptions on discriminability would be mitigated. Future research efforts should manipulate different timings of the verbal descriptions, including a more protracted time course (Mickes, 2016) so that evidence for the optimal time points could be determined.

The CAC results are a matter of importance for judges and jurors who make decisions about culpability (Mickes, 2015; Mickes, 2016). On this issue, the message of our research likely differs from what has thought to be true of the effect of verbal overshadowing. More specifically, our results suggest that, regardless of whether a verbal description was provided, the reliability of an ID made from a lineup or a showup was comparable. Moreover, high-confidence IDs from a lineup were quite accurate in both conditions (greater than 90% correct), whereas high-confidence IDs made from a showup were less accurate in both conditions. The fact that identifications made with high confidence are associated with lower PPV when memory is tested on showups than lineups should signal to judges and jurors that those identifications may be less trustworthy and thus should be taken with caution.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., & Birt, A. R., . . .Zwaan, R.A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556-579.

Brainerd, C. J. & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10,* 3-47.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relation in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11-30.

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 739-755.

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30,* 113-125.

Fortin, N. J., Wright, S. P., & Eichenbaum, H. (2004). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature, 9,* 188-191.

Goodsell, C. A., Wetmore, S. A., Neuschatz, J. S., & Gronlund, S. D. (2013). Showups vs. lineups: A review of two identification techniques. In B. Cutler (Ed.), *Reform of eyewitness identification procedures* (pp.45-64). Washington, DC: American Psychological Association.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23,* 3-10.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-

accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1304-1316.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*, 603-616.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93-102.

Mickes, L. (2016). The effects of verbal descriptions on eyewitness memory: Implications for the real-world. *Journal of Applied Research in Memory and Cognition, 5,* 270-276.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous vs. sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361-376.

Mickes, L. & Wixted, J. T. (2015). On the applied implications of the "Verbal Overshadowing Effect". *Perspectives on Psychological Science, 10,* 400-403.

Nguyen, T. B., Pezdek, K. & Wixted, J. T. (in press). Evidence for a Confidence-Accuracy Relationship in Memory for Same- and Cross-Race Faces. *Quarterly Journal of Experimental Psychology*.

Palmer, M., Brewer, N., Weber, N. & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55-71.

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making, 7,* 332-359.

Rotello, C. M., Heit, E., & Dube, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin and Review, 22,* 944-954.

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337–347.

Schooler, J. W. (1998). The distinctions of false and fuzzy memories. *Journal of Experimental Child Psychology*, 71(2), 130-143.

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

Seale-Carlisle, T. M. & Mickes, L. (2016). US lineups outperform UK lineups. *Royal Society Open Science.* DOI: 10.1098/rsos.160300

Smith, H. M. J., & Flowe, H. D. (2015). ROC analysis of the verbal overshadowing effect: Testing the effect of verbalisation on memory sensitivity. *Applied Cognitive Psychology, 29,* 159-168.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 27,* 523-540.

Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The hippocampus dupports both the recollection and the familiarity components of recognition memory. *Neuron, 49,* 459-466.

Wetmore, S., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition, 4,* 4-18.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121,* 262-276.

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence-accuracy relationship. *Journal of Applied Research in Memory and Cognition, 5,* 192-203.

Table 1. Frequencies of suspect IDs, filler IDs, and no IDs for target-absent and target-present lineups for all levels of confidence in the control and verbal conditions in Experiments 1-4.

	Control					Verbal				
	Target-present			Target-absent		Target-present			Target-absent	
Confidence	Suspect IDs	Filler IDs	No IDs	Filler IDs	No IDs	Suspect IDs	Filler IDs	No IDs	Filler IDs	No IDs
					Experiment 1					
1	3	2		5		1	0		2	
2	1	4		11		3	7		5	
3	10	6		22		5	10		13	
4	17	15	34	35	60	18	12	41	20	77
5	19	13		35		24	14		32	
6	21	9		14		31	5		14	
7	13	4		6		16	3		5	
					Experiment 2					
1	0	2		1		0	1		1	
2	3	3		4		2	5		5	
3	7	5		13		5	8		8	
4	21	10	41	29	86	12	10	87	16	108
5	47	10		33		29	13		29	
6	25	3		10		30	4		14	
7	19	0		3		2	2		4	

	Target-present		Target-absent		Target-present		Target-absent	
				Experiment 3				
	Suspect IDs	No IDs	Suspect IDs	No IDs	Suspect IDs	No IDs	Suspect IDs	No IDs
0%	0		1		0		1	
10%	0		0		1		0	
20%	1		0		0		1	
30%	3		0		2		1	
40%	6		3		7		2	
50%	15	114	7	214	9	120	5	240
60%	25		22		18		9	
70%	51		23		42		15	
80%	47		15		39		8	
90%	30		9		20		4	
100%	36		6		18		7	
				Experiment4				
0%	1		0		0		0	
10%	1		0		0		0	
20%	1		0		1		1	
30%	5		1		0		0	
40%	8		4		4		4	
50%	7	100	10	226	5	149	9	266
60%	23		11		22		11	
70%	52		19		25		11	
80%	54		21		19		14	
90%	36		4		22		5	
100%	23		6		14		1	

Figure 1. Procedural order of the original Experiments 4 (**A**) and 1 (**B**) (Schooler & Engstler-Schooler, 1990); Alogna et al. (2014) RRR1 (**A**) and RRR2 (**B**); and the current Experiments 1 and 3 (**A**) and Experiments 2 and 4 (**B**). (Diagram adapted from Mickes, 2016)

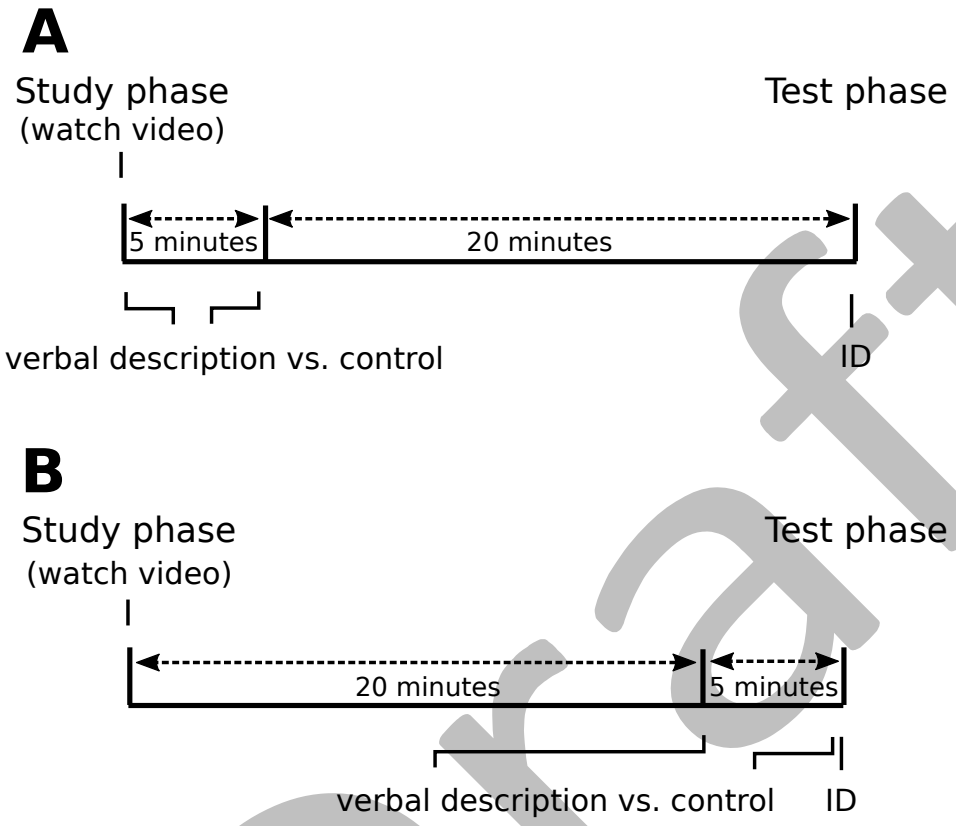


Figure 2. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 1. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.

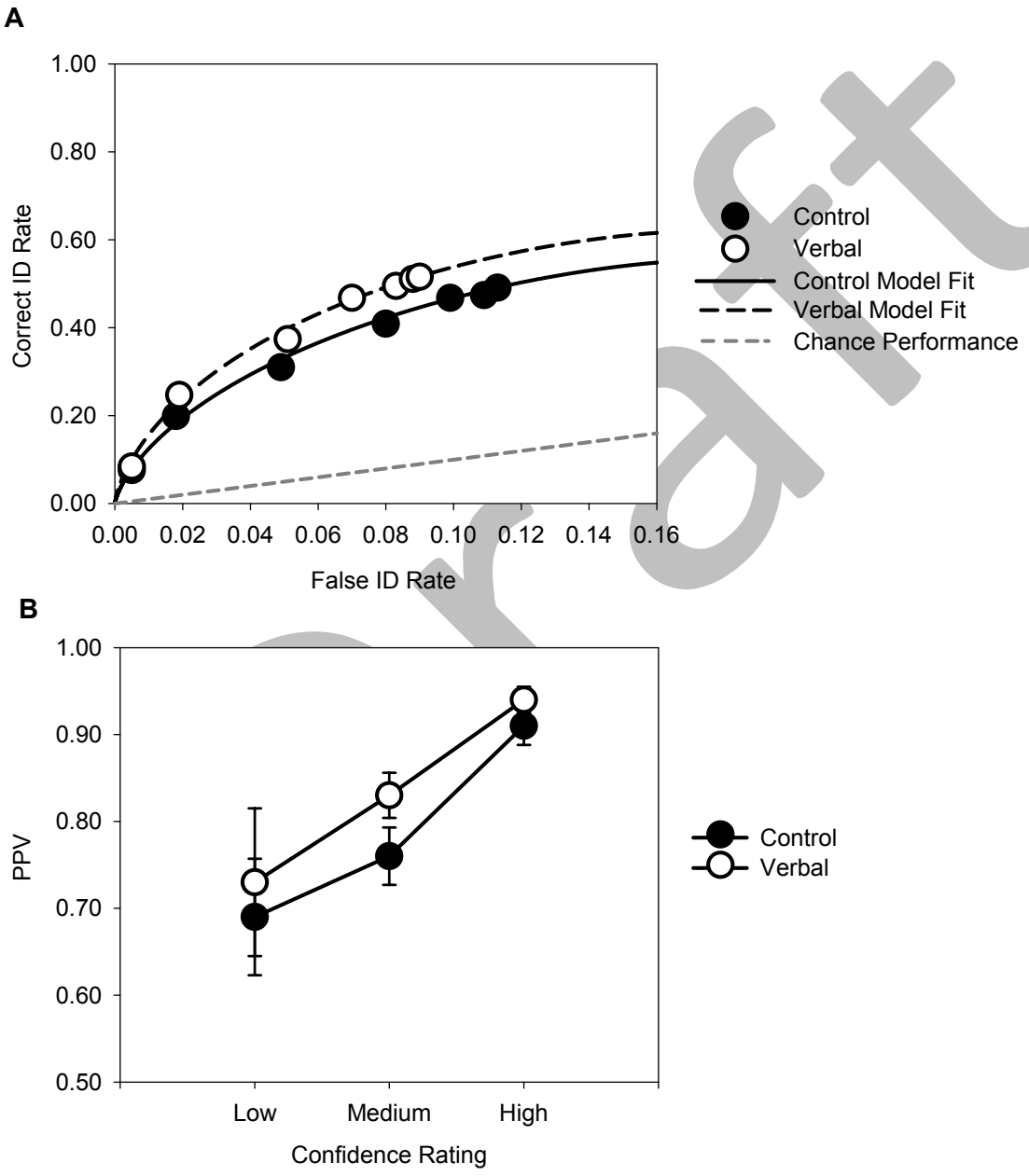


Figure 3. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 2. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.

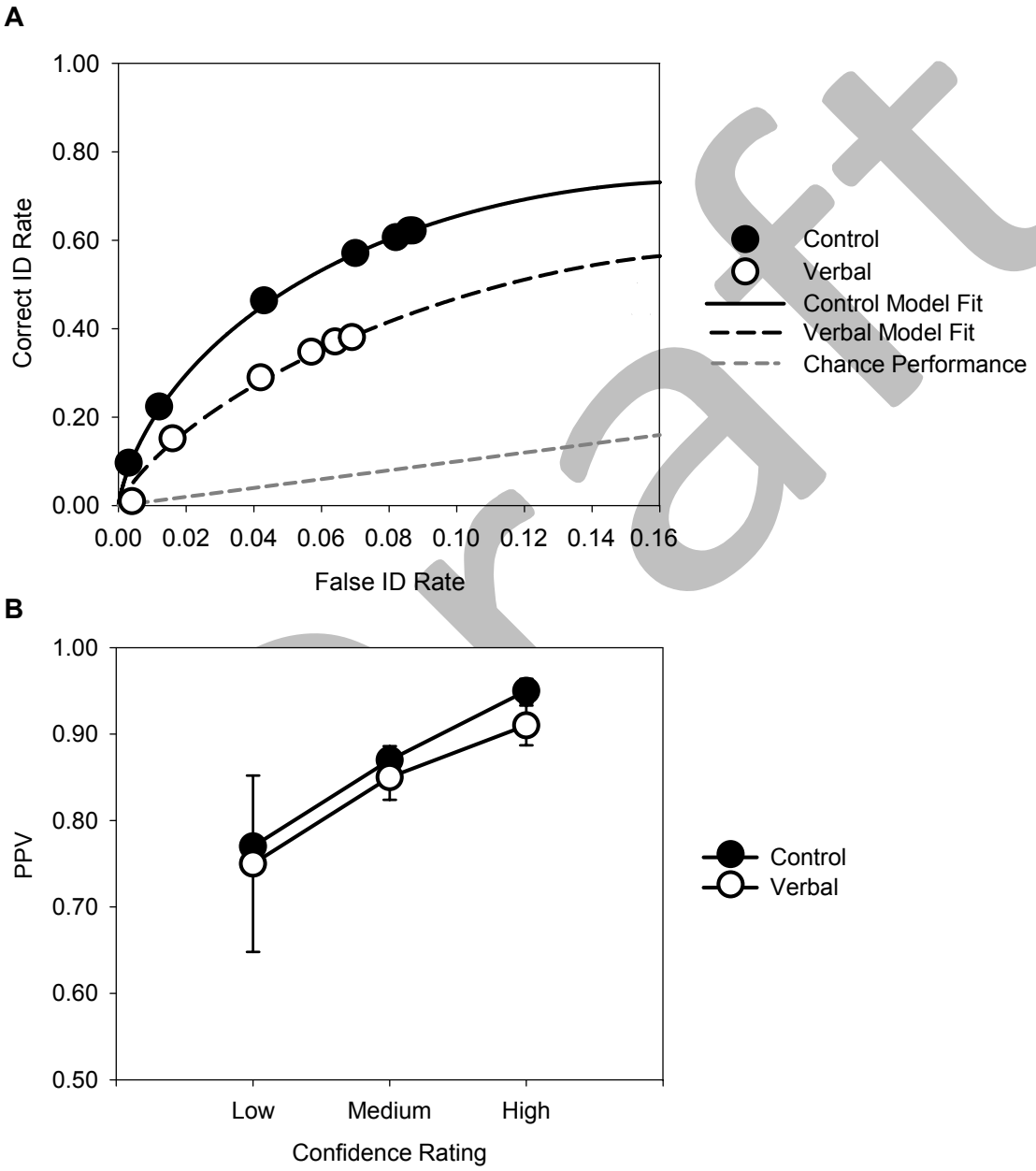


Figure 4. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 3. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.

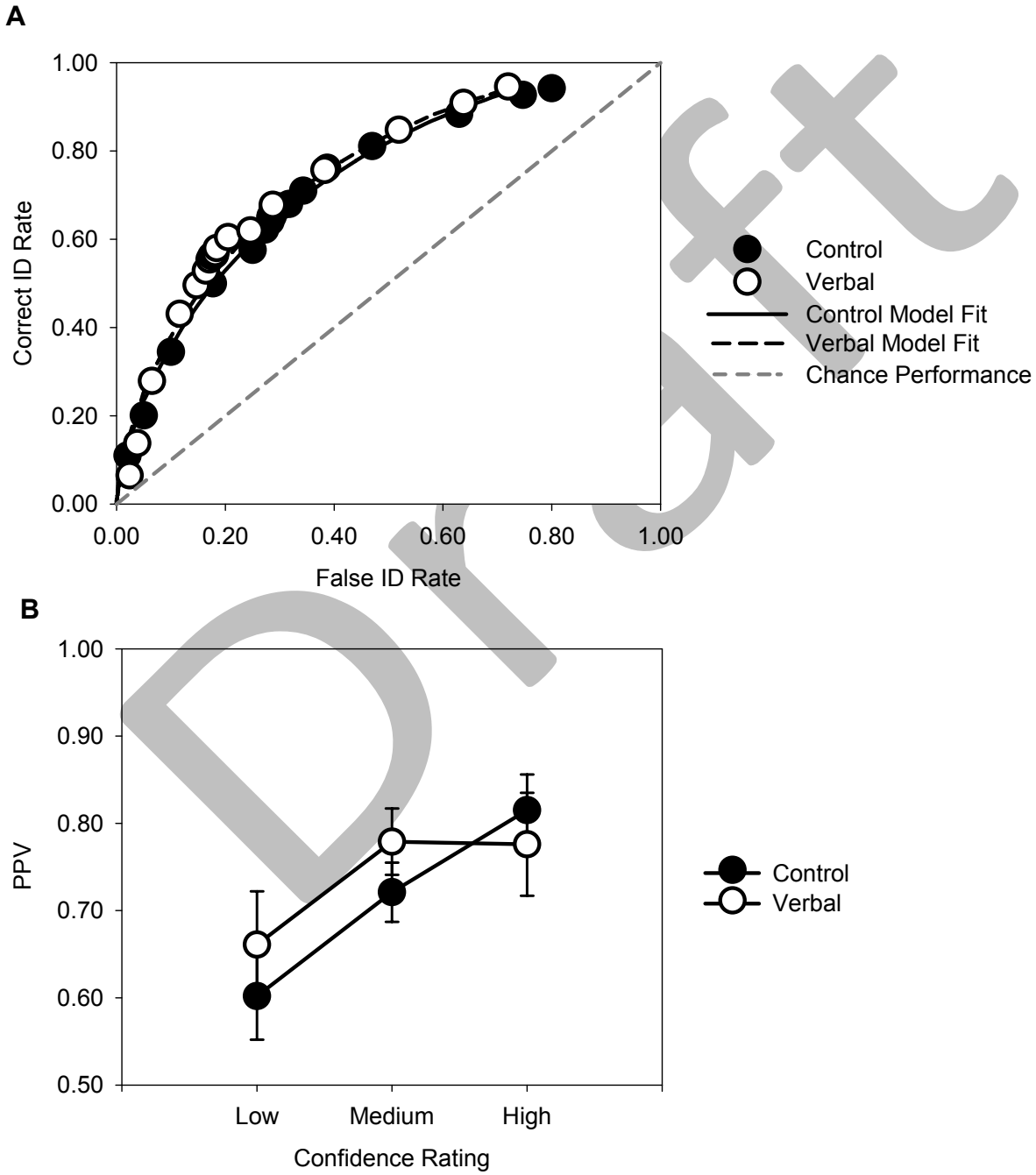
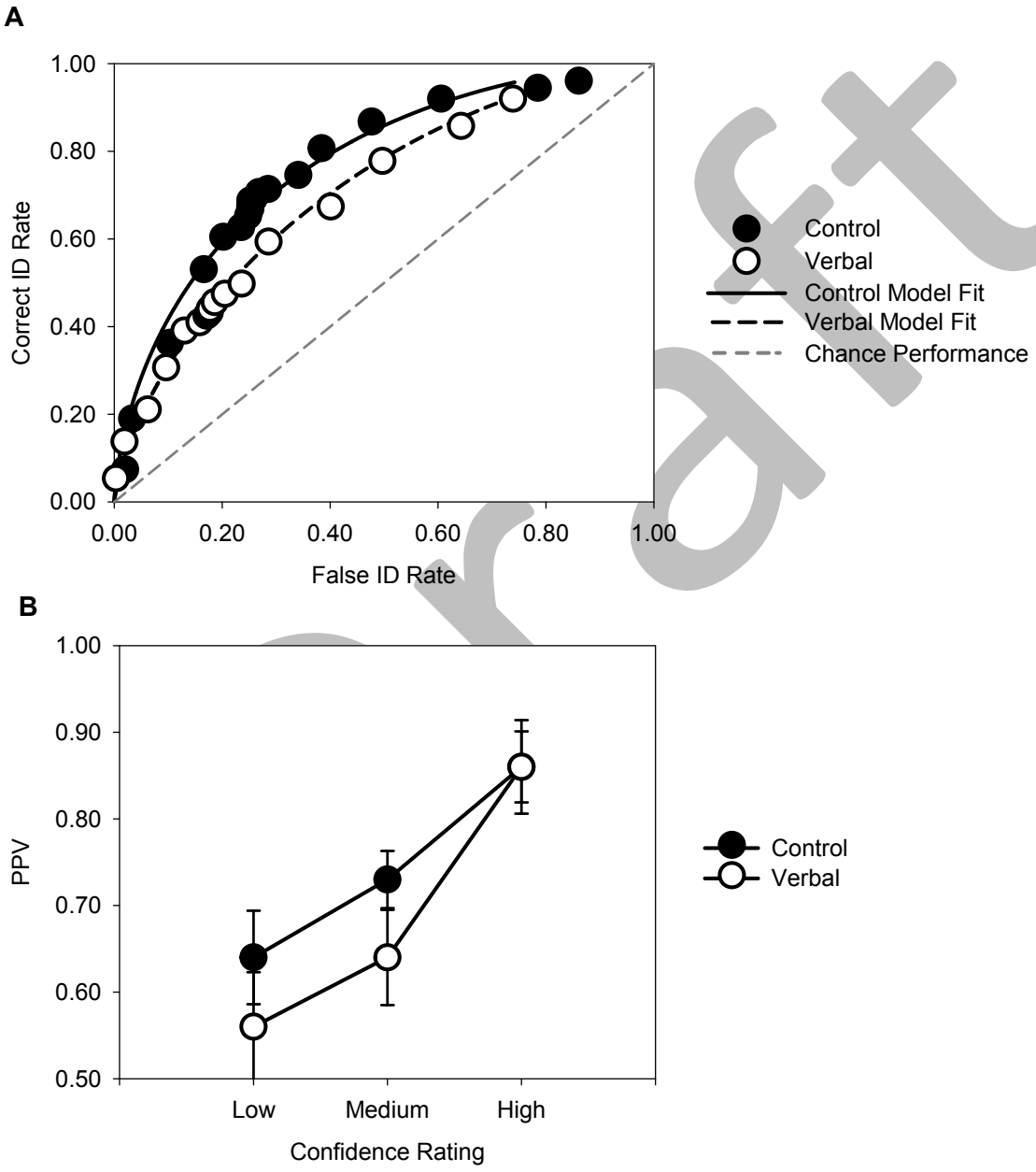


Figure 5. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 4. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.



Appendix

Diagnostic			Non-Diagnostic		
Words	Immediate	Delayed	Words	Immediate	Delayed
chin	76	65	white	491	517
jaw	76	67	male	330	323
cheek	127	89	age	266	295
brow	560	531	brown	449	422
forehead	51	38	black	586	526
eye	673	623	moustache	59	140
oval	33	15	dark	467	555
round	201	186	weight	30	39
wavy	116	86	build	71	94
point	79	53	height	159	161

Note. Different participants used different adjectives. For example, because “chin” and “jaw” were mentioned meant that there was something notable about them that was more diagnostic (e.g., “pointy chin and chiseled jaw”) than ethnicity (“White”) and gender (“male”).