Applied Cognitive Psychology, *Appl. Cognit. Psychol.* **29**: 159–168 (2015)
Published online 30 December 2014 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.3096

ROC Analysis of the Verbal Overshadowing Effect: Testing the Effect of Verbalisation on Memory Sensitivity

HARRIET M. J. SMITH[†] and HEATHER D. FLOWE*

School of Psychology, University of Leicester, Leicester, UK

Summary: This study investigated the role of memory sensitivity versus recognition criterion in the verbal overshadowing effect (VOE). Lineup recognition data were analysed using receiver operating characteristic analysis to separate the effects of verbalisation on memory sensitivity from criterion placement. Participants watched a short crime video, described the perpetrator's facial features and then attempted a lineup identification. Description instructions were varied between participants. There was a standard (free report), forced (report everything) and warning (report accurate information) condition. Control participants did not describe the perpetrator. Memory sensitivity was greater in the control compared with the standard condition. Memory sensitivity was also greater in the warning condition than in the forced and standard conditions. Memory sensitivity did not differ across the forced and standard-description conditions, although a more conservative lineup decision standard was employed in the forced condition. These results, along with qualitative analyses of descriptions, support both retrieval-based and criterion-based explanations of the VOE. Copyright © 2014 John Wiley & Sons, Ltd.

Following a crime, witnesses are likely to be asked questions about the appearance of the perpetrator (TWGEE, 1999). These descriptions may allow law enforcement officers to apprehend suspects fitting witnesses' descriptions, rule out other suspects and construct fair lineups. Obtaining a detailed and accurate verbal description of the perpetrator can be vital in securing a conviction. Some psychological research suggests that in prompting witnesses to describe a perpetrator, the police may be undermining the diagnostic value of lineup identification evidence (e.g. Schooler & Engstler-Schooler, 1990). The act of describing the perpetrator could subsequently make witnesses less likely to identify the guilty suspect from a lineup.

The verbal overshadowing effect (VOE) refers to the tendency for verbal descriptions to impair subsequent recognition performance. In Schooler and Engstler-Schooler's (1990) seminal study, participants watched a 30-second video of a robbery and then attempted to identify the perpetrator on an eight-person target-present lineup. Participants who verbally described the perpetrator's face prior to the lineup were significantly less accurate than those who did not. The effect has been replicated (e.g. Brown & Lloyd-Jones, 2003; Dodson, Johnson & Schooler, 1997; Fallshore & Schooler, 1995; Westerman & Larsen, 1997). Meissner and Brigham's (2001) meta-analysis indicated that verbalisation had a small but significant negative effect on face identification accuracy. Participants were 1.27 times more likely to be inaccurate on target-present lineups. However, the magnitude and direction of the effect vary across studies. Early studies uncovered lineup performance disruptions of around 50% following verbal descriptions (Schooler & Engstler-Schooler, 1990). However, the first replication attempt detected an effect size that was 30% lower (Fallshore & Schooler, 1995). There have also been numerous failures to replicate the VOE (e.g. Clifford, 2003; Memon & Bartlett, 2002; Yu & Geiselman, 1993). Some studies have even observed a facilitating effect of verbal descriptions on face recognition (e.g. Itoh, 2005). Prompted in part by conflicting findings, different theoretical explanations have been put forth to account for the VOE.

* Correspondence to: Heather D. Flowe, School of Psychology, University of Leicester, 106 New Walk, Leicester, UK, LE1 7EA.
E-mail: hf49@le.ac.uk

[†]Harriet M. J. Smith is now at Nottingham Trent University, UK.

CONTENT ACCOUNTS: THE EFFECT OF VERBALISATION ON MEMORY SENSITIVITY

The VOE may occur because the content of verbalisation affects the quality and/or the accessibility of the original visual memory trace. The recoding interference account is the earliest account of the VOE. It proposes that translating hard-to-describe memories of faces into words involves non-veridical recall. The verbal memory competes with the original visual memory during the visual recognition test and reduces recognition accuracy (Schooler & Engstler-Schooler, 1990). Several lines of work provide converging evidence in support of the recoding interference account (e.g. Melcher & Schooler, 1996; Ryan & Schooler, 1998).

Another content account, the retrieval-based interference (RBI) account (Meissner, Brigham & Kelley, 2001), also focuses on how verbalisation may affect the underlying memory trace. This account focuses on how description instruction impacts the memory trace. In particular, when description instructions encourage participants to utilise a liberal criterion in generating the description, the description is more likely to contain errors. These errors, in turn, affect the accuracy of the memory representation. Finger and Pezdek (1999) compared the effect of elaborative forensic description instructions based on the Cognitive Interview (Geiselman et al., 1984) with those from a standard police interview. Participants in the elaborative description condition provided more correct and incorrect facial descriptors and were less accurate at lineup (Finger & Pezdek, 1999). Additionally, Meissner et al. (2001) found that under forced description instructions, participants produced more inaccurate descriptions and performed less accurately at lineup compared with participants who wrote descriptions in the free recall or warning condition. In the forced condition,

participants were instructed to report everything they could remember, even if guessing, whereas in the warning condition, they were instructed to strive for accuracy and only to report information they were confident in. Thus, taken together, these findings indicate that verbalisation can impact the quality of the memory representation.

TRANSFER-INAPPROPRIATE PROCESSING SHIFT

Schooler (see Schooler, 2002, for a review) has since put forward an alternative account of the VOE, arguing that the effect is attributable to a transfer-inappropriate processing shift (TIPS). Accurate face recognition is facilitated by configural processing (Diamond & Carey, 1986). In contrast, verbal descriptions of faces focus on the retrieval of a sequence of features described at the local level (Wells & Turtle, 1988). This mismatch in processing during encoding and retrieval inhibits subsequent lineup performance because the cognitive processes that support accurate recognition are temporarily disrupted. Unlike the explanation put forward by Meissner et al. (2001), TIPS emphasises quantity over quality of verbal descriptions. Schooler (2002) argues that as global rather than local processing is our default processing style (Kimchi, 1992), any shift to local features promoted by verbalisation will be temporary. In line with TIPS, some research has found that the VOE is indeed temporary (Finger & Pezdek, 1999). Additionally, TIPS can account for the finding that verbalisation disrupts recognition of other faces, not just the one that was described (Dodson et al., 1997), as well as findings on the effects of verbalisation on own-race versus other-race face recognition (Schooler, Fiore, & Brandimonte, 1997).

CRITERION SHIFT ACCOUNT

Clare and Lewandowsky (2004) proposed that the VOE occurs because verbalisation primarily affects the decision strategy, or criterion, that participants adopt at the lineup test. According to this account, participants find it difficult to verbally describe facial stimuli and so come to feel uncertain about the accuracy and/or completeness of their description. This makes them doubt the strength of their memory. Consequently, participants who describe the face, compared with those who do not, adopt a more conservative decision standard at lineup. They are less likely to identify any face, including the target. As such, under the criterion shift account, both the hit rate and the false-alarm rate will be lower following verbalisation.

Clare and Lewandowsky (2004) tested their hypothesis by varying whether or not participants gave a description of the perpetrator, and if so, whether the description was featural versus holistic. In the featural condition, participants focused on describing facial features, whereas in the holistic condition, they focused on describing the perpetrator in terms of facial averageness and personality traits. In Experiment 1, participants in verbalisation conditions were less likely than participants in the no-description condition to positively identify a face. This was the case in both target-present and target-absent lineups. Consequently, verbalisation resulted in a lower hit rate and a lower false-alarm rate. (Experiments 2 and 3 did not include a target-absent lineup condition; hence, the results are not discussed here.) These results were interpreted as supporting a criterion shift account of the VOE. However, Clare and Lewandowsky pointed out that their results do not rule out RBI as an explanation of the VOE because their study only used standard-description instructions rather than elaborative or forced description instructions.

Meissner (2002) more directly tested lineup criterion-based and RBI accounts because he included a target-absent lineup and several description instruction conditions designed to affect the accuracy of the description given. The study included a no-description control condition, and three verbalisation conditions (forced, warning and standard). Results were not in line with the criterion-based explanation of the VOE. There was no indication of a conservative lineup criterion shift following any kind of verbalisation. Consistent with the RBI explanation, participants in the forced description condition demonstrated both higher false-alarm rates and lower hit rates than participants in both control and warning recall conditions. These findings suggest that verbalisation impacts memory quality and subsequent recognition accuracy rather than the decision criterion employed at test.

Finally, it should be noted that there is further evidence suggesting that verbalisation accuracy can affect criterion placement. Sauerland, Holub and Sporer (2008) investigated the effect of verbal descriptions on identification accuracy with a retention interval of 1 week between description and identification. Participants watched a video of a crime and then wrote a description of the perpetrator. One week later, they attempted identification of the perpetrator from a target-present or target-absent lineup. Half the participants re-read their description just before seeing the lineup, whereas the other half did not. Although no VOE was observed, a lineup criterion shift was exhibited by re-readers. This group of participants was less likely to make a positive identification from the lineup. Both hit rates and false identification rates were lower compared with those in standard-description and no-description conditions. In the re-reader group, there was also a relationship between description accuracy and choosing. The number of incorrect details reported negatively correlated with choosing any face from the lineup. These results suggest that underlying knowledge about the accuracy of one's descriptions influences choosing behaviour (Sauerland, Holub & Sporer, 2008).

RECEIVER OPERATING CHARACTERISTIC ANALYSIS OF VERBAL OVERSHADOWING EFFECT

One outstanding issue in need of further investigation is whether verbalisation alters memory sensitivity, or the ability to detect the target from the fillers in a lineup. Clare and Lewandowsky's (2004) findings seem to suggest that verbalisation does not impact memory sensitivity but rather changes the decision process. The RBI account, on the other hand, proposes that the content of verbalisation interferes

with (or changes) the underlying memory trace for the to-be-remembered face. Under the TIPS account, verbalisation could impact memory sensitivity or criterion placement (Chin & Schooler, 2008).

On the one hand, Clare and Lewandowsky's (2004) Experiment 1 data may seem to clearly indicate that participants' response criteria at lineup were shifted to a more conservative level as a consequence of verbalisation. Here, however, we argue that their data are also consistent with a memory sensitivity account. In particular, we will use receiver operating characteristic (ROC) analysis to examine the VOE because it allows for the separation of response bias and memory sensitivity in the analysis of recognition data.

Recent eyewitness research has employed ROC analysis to investigate memory processes in lineups (e.g. Gronlund et al., 2012; Mickes, Flowe & Wixted, 2012; Wixted & Mickes, 2012). When a change in testing procedure causes the hit rate and the false-alarm rate to decrease, this could mean that either memory sensitivity or decision criterion placement is varying across procedures. In such cases, ROC analysis is necessary because it separates memory sensitivity from response bias (see Gronlund, Wixted, & Mickes, 2014, for further information about how to conduct a confidence-based analysis of lineup data). We will now illustrate how ROC analysis may be applied to determine whether the VOE occurs because verbalisation impacts memory sensitivity versus criterion placement.

The top and bottom panels of Figure 1 illustrate a criterion shift account (top panel) versus memory sensitivity account (bottom panel) of the VOE, using the hit and false-alarm rates reported by Clare and Lewandowsky (2004), Experiment 1 (see Tables 1 and 2 in their report). The symbols in Figure 1 depict the hit rates (no description: 80%, description: 63%) and false-alarm rates (no description: 77%, description: 48%) they obtained. Note that they tested the criterion hypothesis using the overall false-alarm rate, which included both innocent suspect and filler identifications. For consistency with their report, the false-alarm rates depicted in Figure 1 are also based on innocent suspect and filler identifications. Additionally, in illustrating ROCs for the description condition, we collapsed across the holistic and featural description conditions reported in their paper. Clare and Lewandowsky also collapsed across them in their analysis, because the hit and false-alarm rates obtained did not statistically differ across the conditions.

In the top panel of Figure 1, which illustrates the criterion account, the ROC curves for the no-description and description conditions were fit so that they almost overlap. The key difference between the conditions is that the curve for the description condition is shifted leftward along the *x*-axis relative to the no-description condition, illustrating what the ROCs might look like if a more conservative decision standard at lineup was applied on average in the description condition relative to the no-description condition. A memory sensitivity account is illustrated in the bottom panel of Figure 1; the ROC curves were drawn so that the curve for the no-description condition was distinct from the curve for the description condition. Here, in comparing the two curves, it is apparent that the hit rate is lower and the false-alarm rate is higher at every point along the curves. Therefore, this

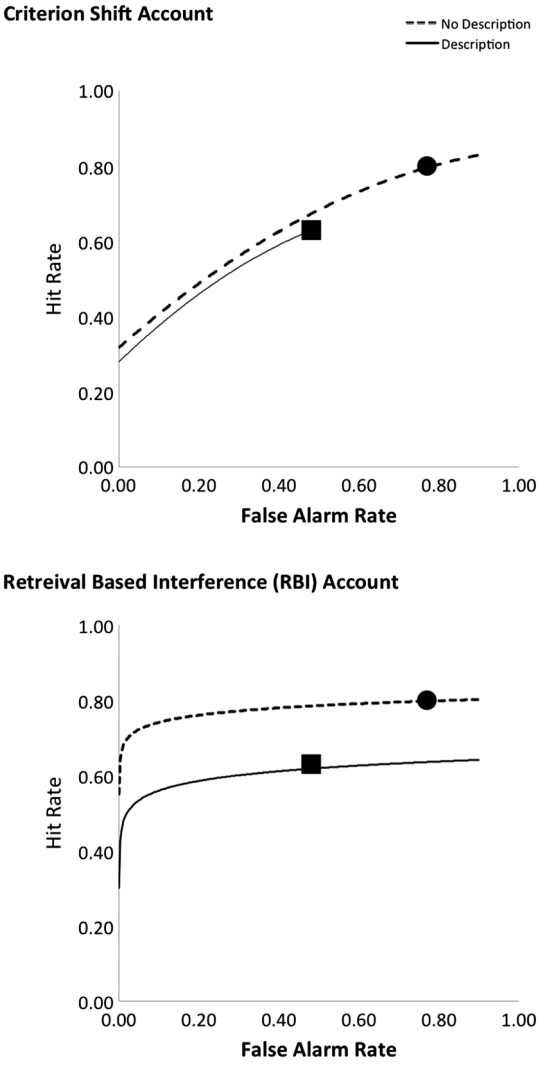

Figure 1. Criterion shift and RBI interpretations of the hit and false-alarm (i.e. innocent suspect and filler identifications) data reported by Clare and Lewandowsky (2004, Experiment 1). The symbols in the figure are the hit and false-alarm rates that they reported for the no-description (i.e. circle symbol) and description (i.e. square symbol) conditions. Clare and Lewandowsky employed two types of description conditions (i.e. featural and holistic), but the results did not differ across conditions; hence, they collapsed across them in their analysis. Therefore, we also collapsed across these conditions to depict the hit and false-alarm rates when a description was given. The top panel of the figure illustrates the criterion shift account of the data, whereby the receiver operating characteristic (ROC) curves for the no-description and description condition almost overlap. Here, the key difference across conditions is that on average participants in the description condition set a higher decision criterion (i.e. they responded more conservatively); hence, they had a lower false-alarm rate compared with those in the no-description condition. Thus, the curve for the description condition is shifted leftward on the *x*-axis relative to the no-description condition. The bottom panel depicts the RBI account, whereby the no-description condition falls on a higher ROC curve than the description condition, illustrating that memory sensitivity was greater on average when no description was given.

illustrates how the ROC curves might look if memory sensitivity was greater when no description was given compared with when one was given. In comparing the top and bottom panels of Figure 1, it is apparent that the data from Clare and Lewandowsky (2004) could be interpreted as consistent with either a criterion or memory sensitivity account of the VOE.

Table 1. Distribution of identification responses across the description conditions on target-present (TP) and target-absent (TA) lineups

	TP	TA
No description ($n = 88$ TP, 83 TA)		
Target	0.63	
Filler	0.15	0.57
Reject	0.23	0.42
Standard ($n = 61$ TP, 65 TA)		
Target	0.56	
Filler	0.28	0.69
Reject	0.16	0.31
Warning ($n = 77$ TP, 73 TA)		
Target	0.53	
Filler	0.16	0.32
Reject	0.31	0.68
Forced ($n = 75$ TP, 71 TA)		
Target	0.41	
Filler	0.13	0.35
Reject	0.45	0.65

Table 2. Descriptive statistics (mean, standard error of the mean (SEM) and 95% confidence interval (CI)) from the qualitative analysis of the descriptions across the standard, forced and warning description conditions

Dependent variable	Description condition	Mean	SEM	95% CI Lower	95% CI Upper
Number of descriptors	Standard	2.67	0.31	2.07	3.28
	Forced	9.46	0.28	8.90	10.02
	Warning	5.53	0.28	4.97	6.08
Correct number of details	Standard	1.17	0.15	0.87	1.46
	Forced	3.70	0.14	3.43	3.97
	Warning	2.67	0.14	2.40	2.94
Incorrect number of details	Standard	0.28	0.06	0.16	0.40
	Forced	0.62	0.06	0.51	0.73
	Warning	0.28	0.05	0.17	0.39
Subjective number of details	Standard	1.23	0.22	0.79	1.67
	Forced	5.14	0.21	4.74	5.55
	Warning	2.58	0.20	2.18	2.98
Proportion accurate	Standard	0.47	0.02	0.42	0.52
	Forced	0.42	0.02	0.38	0.47
	Warning	0.51	0.02	0.46	0.55

Interestingly, the Meissner (2002) data are consistent with a change in memory sensitivity rather than a change in criterion placement following verbalisation, as the false-alarm rate was lower and the hit rate was higher in the no-description condition than in the description condition.

AIMS OF THE PRESENT STUDY

Given the conflicting nature of the findings reported by Clare and Lewandowsky (2004) and Meissner (2002), it seems prudent to employ ROC analysis to more directly assess whether verbalisation negatively impacts memory sensitivity. Note that content accounts predict that verbalisation will impact memory sensitivity. The TIPS account proposes that a processing shift can affect either the memory trace or the decision criterion employed at test (Chin & Schooler, 2008).

We included warning, forced and standard-description conditions, as well as a no-description control condition. If verbalisation negatively impacts the memory trace and/or its accessibility, memory sensitivity should be greater when no description is given than when a standard description is given. Additionally, the criterion employed in generating the description should also impact memory sensitivity. Participants should write descriptions that contain fewer errors if they are warned to only include information that they believe is accurate (warning, or high-criterion instructions) compared with when participants are forced to include everything they can remember, even if they are guessing (forced, or low-criterion instructions). RBI explanations propose that memory disruption should be greater when the description contains more errors. Memory sensitivity should be greater in the warning compared with the standard (free report) and forced conditions, and memory sensitivity should be greater in the standard compared with the forced condition.

We also analysed the content of the descriptions to assess whether description quality varied as a function of verbalisation condition. If we successfully affected the decision criterion participants employed in writing the description, forced descriptions should contain the greatest number of correct, incorrect and subjective details, followed by descriptions in the standard and warning conditions. We also tested the prediction that identification accuracy would be less accurate when the number of errors in the report was relatively high, which would be expected if verbalisation impacts memory sensitivity.

METHOD

Participants

A total of 593 participants (42% women) were recruited online for the study. The majority of participants reported being White (63%); other backgrounds reported included East Asian (14%), South Asian (11%), Black (7%) and other ethnic/racial backgrounds (5%). Participants were between 18 and 76 years old ($M = 31.91$ years, $SD = 12.71$ years).

Materials

Video

Participants viewed a 24-second video featuring a 22-year-old White American man stealing a laptop from an empty office. His face is clearly visible as he walks out of the office with the laptop. He pauses at the door to look both ways before exiting.

Lineup construction

Defining features of the perpetrator such as race, age (20–30 years), eye colour and hair colour were entered into the Florida Department of Corrections Offender Network database (http://www.dc.state.fl.us/AppCommon/). In total, 22 pictures were selected from the resulting matches and used as foils to construct four different

lineups. Both target and foils had neutral facial expressions. They did not have any distinctive features such as piercings or tattoos. All pictures were edited to show only the face and neck. There were two 6-person target-present lineups, with the target appearing in either Position 2 or 5 and two 6-person target-absent lineups. Target-absent lineups did not feature a designated innocent suspect. Different fillers were used for each of the four lineups.

Pilot testing

Research has shown that the VOE may be dependent on test-set similarity (Kitagami, Sato & Yoshikawa, 2002), making it necessary to construct lineups of verbally similar faces, as in Schooler and Engstler-Schooler (1990). Pilot testing was conducted to test lineup fairness.

Four participants watched the video of the crime and then described the perpetrator. A separate group of participants ($N = 28$) served as mock witnesses. Mock witnesses read one of the four descriptions, which was randomly assigned, and then viewed one of the lineups (presented simultaneously). They attempted to identify the perpetrator in the lineup on the basis of only the description. This procedure was repeated until all four lineups were evaluated by the mock witness. Tredoux's E was calculated for each of the four lineups to measure the number of lineup members who matched the description of the perpetrator (Malpass, 1981; Tredoux, 1998). Average Tredoux's E for the four lineups was 3.93. This corresponds well with archival study estimates (e.g. Valentine & Heaton, 1999) and previous laboratory studies using 'fair' lineups (Gronlund, Carlson, Dailey, & Goodsell, 2009).

Design and procedure

The University of Leicester's Ethics Committee granted ethical approval for the study. The experiment employed a 4×2 independent group factorial design. The factors were post-encoding instructions (no description, standard description, warning description or forced description) and lineup type (target present or target absent).

Participants were randomly allocated to conditions. They read an information sheet, completed a consent form and recorded their age, gender and ethnicity. The study employed an incidental test of memory. Participants were told only that the study was concerned with person perception. Although they were informed that the study would involve a written task, participants were unaware that they might have to describe the target. Participants watched the video and then completed a 5-minute mathematical filler task.

Participants in the three description conditions had 5 minutes to describe the perpetrator, in line with instructions in other verbal overshadowing studies (e.g. Itoh, 2005; Perfect, Hunt & Harris, 2002; Schooler & Engstler-Schooler). Instructions were based on those used by Meissner (2002) and Meissner et al. (2001), with the added instruction to focus on featural aspects of the face (Finger & Pezdek, 1999). We focused on featural descriptions because research has found that police ask eyewitnesses to provide information about features (Fahsing, Ask, & Granhag, 2004) to aid the construction of fair lineups (Lindsay, Martin & Webber, 1994). In the standard condition, participants were given the following instructions:

> In the box below, please describe the face you saw in the video. 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 focus on FACIAL FEATURES. Write about the shape and size of the eyes, eyebrows, nose, ears, mouth, chin etc.

In addition to these instructions, participants in the warning condition were instructed:

> Prior research has demonstrated the importance of striving for ACCURACY and reporting only that which you are CERTAIN you remember. You should attempt to give the MOST ACCURATE DESCRIPTION OF THE FACE POSSIBLE. Be sure to report only those details that you are confident of, and DO NOT ATTEMPT TO GUESS at any particular feature.

In both the standard and warning conditions, participants were told to stay on the description page for 5 minutes but that they did not have to continue writing for the full 5 minutes. In the forced description condition, participants were instructed as per the standard condition, as well as being given the following supplementary instructions:

> Prior research has demonstrated the importance of reporting EVERYTHING that you can remember. Try not to leave out ANY details about the face even if you think they are not important. You should attempt to give the FULLEST DESCRIPTION OF THE FACE POSSIBLE, even if you start to feel that you are guessing. You should CONTINUE WRITING FOR 5 MINUTES. Please use the whole five minutes to write your description.

In the no-description condition, participants were given 5 minutes to complete a verbal listing task, similar to that used in other studies (e.g. Fallshore & Schooler, 1995; Ryan & Schooler, 1998; Schooler & Engstler-Schooler, 1990). Participants were required to list as many items as possible that would fit into certain categories (e.g. European car manufacturers).

At test, participants saw one of three different lineups. In the simultaneous condition, all six faces were numbered and presented on the same slide, in two rows of three. In serial and sequential conditions, slides of each face were shown one after the other. Each slide was numbered. The serial and sequential conditions differed in two ways. In the former presentation format, participants were informed of how many faces would be presented prior to the lineup, whereas in the latter they were not. Participants only responded after seeing all the faces. In both the simultaneous and serial conditions, participants were required to select the number of the face (1–6) that had appeared in the video after viewing all faces or to select 'perpetrator is not present in the lineup'. In the sequential lineup, participants were unaware how many faces would be presented and were required to select 'yes' or 'no' in response to the question 'is this the perpetrator?' after seeing each face.

Prior to each lineup, participants were told that the target may or may not be present, in accordance with recommendations (TWGEE, 1999). Confidence ratings were gathered on 9-point Likert-style rating scales. Immediately after making a lineup selection, participants were asked, 'How confident are you that you have made the correct decision?' (1 = *not at all confident, I was just guessing*, 9 = *I am certain that I have made the correct decision*). All participants were debriefed after the study.

Both authors coded the descriptions, and disagreements were resolved as they arose. Descriptions were coded for quality: both accuracy and subjectivity. Corresponding with the procedure of Meissner et al. (2001), correct descriptors were those that correctly described the target. These included references to eye colour, hair colour and hair length. Incorrect descriptors did not correctly describe the target, for example, stating that the target had facial scars or wore an earring. Subjective descriptors were more relative or ambiguous, including references to personality or face shape. Descriptors were coded as subjective if, with a photo of the target visible, people could still have argued about the description. This definition of subjective included references to facial feature size, such as, 'he had a large mouth'. The total number of facial descriptors was recorded for each description.

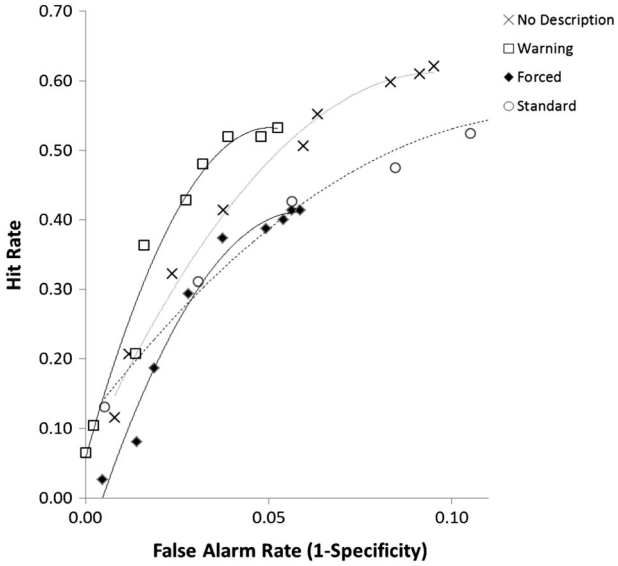

Figure 2. Receiver operating characteristic analysis of hit and false-alarm data from the present study, conditioned on description conditions, including no description, warning, forced and standard. Target-absent lineups did not contain a designated innocent suspect; therefore, false-alarm rates obtained at each confidence level were divided by the total number of faces in the lineup (i.e. six) to obtain an estimate of what the false-alarm rate would have been if an innocent suspect had been designated

RESULTS

Data analysis

We collapsed the data across lineup presentation format. First, previous research has found that the VOE holds across different lineup formats (Meissner, 2002). Second, there were not enough data in each cell (4 description conditions × 9 confidence levels × 3 lineup formats × 2 target conditions) to permit formal analysis. Third, we are not aware of any theoretical reason why verbalisation condition and lineup format would have an interactive effect on memory sensitivity. Lineup format, therefore, will not be discussed further.

Receiver operating characteristic curves were generated for each description condition with confidence rating and identification response data using the following approach. The number of hits in the target-present condition and the number of false-alarms in the target-absent condition were tabulated at each confidence level. The cumulative hit and false-alarm rates were then computed, starting at the highest confidence level and ending with the lowest confidence level. Identification responses by the description condition are presented in Table 1, and ROC results are shown in Figure 2. In Figure 2, the false-alarm rates obtained at each confidence level were divided by the total number of faces in the lineup (i.e. six) to obtain an estimate of what the false-alarm rate would have been if an innocent suspect had been designated, following Mickes et al. (2012).

A partial area under the curve (pAUC) analysis was conducted to compare memory sensitivity across the description conditions. pAUC analysis is appropriate because the false identification rate is typically less than one in lineup research (Clark, Howell, & Davey, 2008). In the present study, the false-alarm rate did not reach 100% in any target-absent condition. pROC, a data analysis tool pack for R (Robin et al., 2011), was used to make the pAUC comparisons. Specificity (1 − FA rate) was set in the analysis using the minimum false-alarm rate obtained in the conditions being analysed. The bootstrap method was used, with the number of replications set to 10 000. Alpha was set to .05 in the analyses. pROC uses the following formula for comparing the two partial areas:

$$D = \frac{\mathrm{AUC1} - \mathrm{AUC2}}{s}$$

where s is the standard deviation of the bootstrap differences and AUC1 and AUC2 are the areas under the curve of the two ROC curves. Hence, D is the difference between the AUCs across two given experimental conditions expressed in standard deviation units, and thus, it may be interpreted as an effect size.

Identification performance

Memory sensitivity in the no-description condition was significantly better than in the standard condition, $D = 2.28$, $p < .05$, ($\mathrm{pAUC}_{\text{no description}} = 0.24$ vs $\mathrm{pAUC}_{\text{standard description}} = 0.16$; $\mathrm{pAUC}_{\text{no description}} - \mathrm{pAUC}_{\text{standard description}} = 0.08$, $SE_{(\text{pAUC no description}-\text{pAUC standard description})} = 0.03$). Thus, the hypothesis that verbalisation negatively impacts the memory trace (or its accessibility) was supported. Additionally, type of description instruction also impacted memory sensitivity. Specifically, in the warning condition, memory sensitivity was significantly better compared with the forced condition, $D = 1.75$, $p < .05$ ($\mathrm{pAUC}_{\text{warning}} = 0.12$ vs $\mathrm{pAUC}_{\text{forced}} = 0.08$; $\mathrm{pAUC}_{\text{warning}} - \mathrm{pAUC}_{\text{forced}} = 0.04$, $SE_{(\text{pAUC warning}-\text{pAUC forced})} = 0.02$) and when compared with the standard condition, $D = 2.45$, $p < .01$ ($\mathrm{pAUC}_{\text{warning}} = 0.12$

vs $pAUC_{standard} = 0.06$; $pAUC_{warning} - pAUC_{forced} = 0.06$, $SE_{(pAUC\ warning - pAUC\ forced)} = 0.02$). Memory sensitivity did not vary across the forced and standard conditions. Given this null result, we turned to the false-alarm rate data to test whether the lineup decision standard in the forced condition was more conservative compared with the standard condition. The false-alarm rate was significantly lower in the forced compared with standard condition when the target was absent (35% vs 69%, respectively), $\chi^2 = 10.67$, $p < .01$, and when the target was present (13% vs 28%), $\chi^2 = 4.47$, $p < .05$.

Description accuracy

Description condition had a significant effect on the number of descriptors given, $F(2, 419) = 134.01$, $p < .001$, $\eta_p^2 = 0.39$. *Post hoc* comparisons using Tukey's honestly significant difference (HSD) test indicated that descriptions were shorter in the warning condition compared with the forced condition and longer in the warning compared with the standard condition (p's $< .0001$). Descriptions were also longer in the forced compared with standard condition ($p < .0001$). Thus, description length was shortest in the standard condition, intermediate in the warning condition and longest in the forced condition.

The mean numbers of correct, incorrect and subjective descriptors by description condition are shown in Table 2. Following Meissner et al. (2001) and Meissner (2002), a multivariate analysis of variance (ANOVA) was employed to analyse the descriptions, with numbers of correct, incorrect and subjective descriptors as the dependent variables and description condition as the independent variable; the result was statistically significant, $F(6, 836) = 37.40$, $p < .001$, $\eta_p^2 = 0.21$. Follow-up univariate ANOVAs were conducted for each dependent variable. Results indicated that description condition had significant effects on the number of correct descriptors reported, $F(2, 419) = 77.09$, $p < .001$, $\eta_p^2 = 0.27$, the number of incorrect descriptors reported, $F(2, 419) = 12.11$, $p < .001$, $\eta_p^2 = 0.05$, and the number of subjective descriptors reported, $F(2, 419) = 89.57$, $p < .001$, $\eta_p^2 = 0.29$. *Post hoc* comparisons using Tukey's HSD test indicated that participants in the warning description condition compared with participants in the forced condition reported significantly fewer correct ($d = 0.54$), incorrect ($d = 0.47$) and subjective ($d = 0.89$) details (p's $< .0001$). Additionally, participants in the warning description condition compared with participants in the standard condition reported significantly more correct ($d = 1.05$) and subjective ($d = 0.74$) details (p's $< .0001$). The number of incorrect details did not significantly vary across the warning and standard conditions. Participants in the forced condition compared with those in the standard condition reported significantly more correct ($d = 1.52$), incorrect ($d = 0.46$) and subjective ($d = 1.47$) details (p's $< .0001$).

The proportion of descriptors that were accurate, which was computed by dividing the number of accurate descriptors by the sum of the total number of correct, incorrect and subjective descriptors, was entered into an ANOVA. Participants who did not provide any correct descriptors were excluded from the analysis. Descriptive results are provided in Table 2. The effect of description condition was statistically significant, $F(2, 400) = 3.48$, $p < .05$, $\eta_p^2 = 0.02$. Tukey's HSD test indicated that participants in the forced condition were significantly less accurate compared with participants in the warning condition ($d = 0.43$, $p < .05$). No other pairwise comparisons were statistically significant.

The relationship between identification accuracy and description accuracy was also examined. Identification accuracy was not significantly associated with the proportion of accurate descriptors within any description condition (forced $r = -.04$, warning $r = -.08$, standard $r = .01$) or overall, when the data were collapsed across description condition ($r = -.03$). Identification accuracy was also not significantly associated with the number of incorrect descriptors reported within any description condition (forced $r = .01$, warning $r = -.06$, standard $r = -.06$) or overall, when the data were collapsed across description condition ($r = -.03$).

DISCUSSION

We employed confidence-based ROC analysis to assess the effects of verbalisation on memory sensitivity versus decision criterion placement, which heretofore has never been performed. We found that memory sensitivity at lineup was poorer for participants who gave a standard description of the perpetrator compared with those who did not give a description. Our findings, therefore, extend and replicate Schooler and Engstler-Schooler's (1990) seminal research. Our results support meta-analytic conclusions that the VOE is a genuine and reliable phenomenon (Meissner & Brigham, 2001).

Following Meissner et al. (2001) and Meissner (2002), we manipulated description instructions to influence the verbal recall criterion participants used when describing the perpetrator. We included this manipulation in order to test whether memory sensitivity varies as a function of instruction condition. Meissner's (2002) findings provide evidence that verbalisation impacts memory sensitivity; the hit rate was higher, and the false-alarm rate was lower in the warning condition compared with the forced condition. We replicated and extended these findings by using ROC analyses, further demonstrating that the quality of verbalisation can impact the underlying memory trace and/or its accessibility. In the warning condition, participants performed more accurately at lineup than those in the forced and standard conditions. This suggests that the quality of the original visual memory trace was less affected in the warning condition compared with the forced and standard conditions. Overall, this pattern of findings is in keeping with previous research on description instruction effects (Finger & Pezdek, 1999; Meissner, 2002; Meissner et al., 2001), as well as research on the effect of warning instructions in the Deese/Roediger–McDermott paradigm. Such instructions have been shown to increase participants' ability to resist false recognition (see Gallo, 2010, for a review). The effect is attributed to the fact that the warning instruction prompts participants to use memory monitoring strategies to resist critical lures (Gallo, Roberts & Seamon, 1997). In the warning condition of the present study, participants may have employed a comparable strategy, perhaps thinking about what the perpetrator did *not* look

like, as well as what he *did* look like. This could have helped them to discriminate between the target and foils at lineup.

Our qualitative analysis of the description data can also be considered in relation to the TIPS and RBI accounts. The TIPS account predicts that lineup identification accuracy decreases as description length increases, because participants shift more from holistic to component-based processing. In the present study, participants in each description condition were instructed to focus on facial features when generating their descriptions. Therefore, longer descriptions should be generally associated with a greater shift to component-based processing. On the one hand, our results might seem consistent with TIPS because participants in the warning condition produced shorter descriptions and performed better at lineup than those in the forced condition. However, participants in the standard condition produced shorter descriptions compared with those in the warning condition, yet their identification performance was poorer compared with those in the warning condition. Thus, TIPS was not fully supported by our findings.

In contrast to TIPS, the RBI account posits that the quality of the underlying visual memory trace is affected by the accuracy rather than the length of the description. Identification performance has been found to be positively associated with description accuracy (Meissner et al. 2001; Meissner, 2002). In line with the RBI account, we found that a greater number of correct descriptors were reported and memory performance was better in the warning condition compared with the forced and standard conditions. Meissner et al. (2001) found that description accuracy was greatest in the warning condition, followed by the standard condition and then the forced. This is consistent with the hypothesis that participants' description criterion is most liberal in the forced condition, intermediate in the standard condition and most conservative in the warning condition. However, in the present study, the overall accuracy of the description was better only in the warning condition compared with the forced condition. Overall accuracy did not differ across the warning and standard-description conditions, despite the fact that lineup performance was better in the warning compared with standard condition. Moreover, in the standard compared with forced condition, fewer correct, incorrect and subjective details were reported, and yet identification performance did not differ across these conditions. There was no correlation between identification accuracy and description accuracy. As such, our description results only partially support RBI. There are some differences between our methodology and that of Meissner et al. (2001) that may account for this discrepancy. The description accuracy data suggest that in the present study, we were perhaps not as successful at manipulating participants' response criterion. The inconsistency, as described earlier, between present description accuracy results and those of Meissner et al. may be due to differences in how descriptions were obtained. Participants in the present study were instructed to continue writing for 5 minutes. Participants of Meissner et al. had a sheet of paper with 25 numbered lines. In the forced condition, they were told to fill in all 25 lines. This procedural difference may have affected the type of descriptions generated in the forced conditions of these two studies. Compared with those of Meissner et al., our participants in the forced condition generated relatively short descriptions ($M = 9.46$ descriptors, 95% confidence interval [8.90, 10.02] descriptors; please see Table 2 of Meissner et al. for comparison) and produced fewer incorrect descriptors. Given these differences, it may not be prudent to necessarily expect our results to replicate those of Meissner et al. All things considered, the present results certainly do not rule out RBI.

We found some evidence supporting the hypothesis that participants monitor the accuracy of their description and then use this information to set their lineup decision criterion (Clare & Lewandowsky, 2004; Sauerland, Holub, & Sporer, 2008). Participants in the forced condition adopted a more conservative lineup decision criterion than those in the standard condition. In the forced condition compared with the standard condition, participants gave longer descriptions containing more correct, incorrect and subjective details. Results therefore suggest that participants in the forced condition estimated the accuracy of their memory to be relatively poor compared with those in the standard condition. This perhaps dissuaded them from choosing any face from the lineup.

Results of the present study have applied forensic relevance. Witness descriptions are important in facilitating arrests. We echo calls made by others for the police to use caution when asking witnesses to describe a criminal perpetrator immediately prior to attempting identification at lineup (e.g. Meissner, 2002). Our results indicate that if the police ask witnesses to provide an exhaustive description of a perpetrator, it might decrease identification accuracy. It may be prudent for the police to caution witnesses against guessing information about the perpetrator's appearance. Instead, they might want to encourage witnesses to report only information about which they are confident. Our study adds to research (Meissner et al., 2001; Meissner, 2002) indicating that memory sensitivity is better when participants are warned against providing incorrect information in their description.

In order to help guide future research in the field, we now turn to limitations of the current study. First, we used a short retention interval between encoding and test. This methodological choice is in line with other VOE studies (Clare & Lewandowsky, 2004; Meissner, 2002, Experiment 1; Meissner et al., 2001). Other studies have found a 'release from the verbal overshadowing effect' after longer retention intervals (Finger & Pezdek, 1999). Having said this, VOEs can persist longer (e.g. Meissner, 2002, Experiment 2). In order to guide legal professionals about the optimal time for gathering perpetrator descriptions, further work should investigate whether the memory sensitivity and criterion placement effects observed here hold at longer retention intervals. Second, our work, like some of that we sought to replicate and extend (e.g. Clare & Lewandowsky, 2004; Finger & Pezdek, 1999; Meissner et al., 2001), employed a single perpetrator. Ideally, research should ensure that the effects obtained generalise to other stimulus materials (Wells & Windschitl, 1999). Having said that, our findings are consistent with previous research. We can think of no theoretical reason why they would not generalise across different types of stimulus materials. We encourage other labs to replicate our ROC findings using other

stimulus materials. Third, although we had no reason to expect the VOE to vary across lineup procedure type (Meissner, 2002), we were unable to formally test this owing to sample size limitations. ROC analyses require large sample sizes to ensure that enough participants select a given confidence rating in both the target-absent and target-present conditions.

In sum, our findings add to a small but new body of research indicating that ROC analysis of lineup data can shed light on memorial processes. We found evidence that verbalisation can impact memory sensitivity. Thus, our data provide compelling evidence that verbalisation affects the quality and/or accessibility of the original visual memory trace. Our results highlight the complex nature of the VOE because we also found evidence of shifts in decision criterion, thus suggesting that participants were monitoring the accuracy of their memory. We hope that other studies adopt the ROC approach when analysing the VOE, thereby further clarifying how verbalisation affects memorial processes.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor and anonymous reviewers, whose helpful comments on earlier drafts have greatly improved this manuscript.

REFERENCES

Brown, C., & Lloyd-Jones, T. (2003). Verbal overshadowing of multiple face and car recognition: Effects of within- versus across-category verbal descriptions. *Applied Cognitive Psychology*, *17*, 183–201. DOI: 10.1002/acp.861

Chin, J. M., & Schooler, J. W. (2008). Why do words hurt? Content, process, and criterion shift accounts of verbal overshadowing. *European Journal of Cognitive Psychology*, *20*, 396–413. DOI: 10.1080/09541440701728623

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 739–755. DOI: 10.1037/0278-7393.30.4.739

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior*, *32*, 187–218. DOI: 10.1007/s10979-006-9082-4

Clifford, B. R. (2003). The verbal overshadowing effect: In search of a chimera. In M. Vanderhallen, G. Verwaeke, P. J. van Koppen, & J. Goethals (Eds.), *Much ado about crime: Chapters on psychology and law* (pp. 151–162). Brussels: Politeia.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*, 107–117. DOI: 10.1037/0096-3445.115.2.107

Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. *Memory & Cognition*, *25*, 129–139. DOI: 10.3758/BF03201107

Fahsing, I. A., Ask, K., & Granhag, P. A. (2004). The man behind the mask: Accuracy and predictors of eyewitness offender descriptions. *Journal of Applied Psychology*, *89*, 722–729. DOI: 10.1037/0021-9010.89.4.722

Fallshore, M., & Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1608–1623. DOI: 10.1037/0278-7393.21.6.1608

Finger K., & Pezdek K. 1999. The effect of cognitive interview on face identification accuracy: Release from verbal overshadowing. *Journal of Applied Psychology*, *84*, 340–348. DOI: 10.1037/0021-9010.84.3.340

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833–848. DOI: 10.3758/MC.38.7.833

Gallo, D. A., Roberts, M. J., & Seamon, J. G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin & Review*, *4*(2), 271–276. DOI: 10.3758/BF03209405

Geiselman, R. E., Fisher, R. P., Firstenberg, I., Hutton, L. A., Sullivan, S. J., Avetissian, I. V., & Prosk, A. L. (1984). Enhancement of eyewitness memory: An empirical evaluation of the cognitive interview. *Journal of Police Sciences and Administration*, *12*, 74–80. DOI: 10.1037/0021-9010.70.2.401

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *15*, 140–152.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221–228. DOI: 10.1016/j.jarmac.2012.09.003

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*, 3–10. DOI: 10.1177/0963721413498891

Itoh, Y. (2005). The facilitating effect of verbalization on the recognition memory of incidentally learned faces. *Applied Cognitive Psychology*, *19*, 421–433. DOI: 10.1002/acp.1069

Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: A critical review. *Psychological Bulletin*, *112*, 24–38. DOI: 10.1037/0033-2909.112.1.24

Kitagami, S., Sato, W., & Yoshikawa, S. (2002). The influence of test-set similarity in verbal overshadowing. *Applied Cognitive Psychology*, *16*, 963–972. DOI: 10.1002/acp.917

Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy. *Law and Human Behavior*, *18*, 527–541. DOI: 10.1007/BF01499172

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior*, *5*, 299–309. DOI: 10.1007/BF01044945

Meissner, C. A. (2002). Applied aspects of the instructional bias effect in verbal overshadowing. *Applied Cognitive Psychology*, *16*, 911–928. DOI: 10.1002/acp.918

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, *15*, 603–616. DOI: 10.1002/acp.728

Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition*, *29*, 176–186. DOI: 10.3758/BF03195751

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, *35*, 231–245. DOI: 10.1006/jmla.1996.0013

Memon, A., & Bartlett, J. (2002). The effects of verbalization on face recognition in young and older adults. *Applied Cognitive Psychology*, *16*, 635–650. DOI: 10.1002/acp.820

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. DOI: 10.1037/a0030609

Perfect, T. J., Hunt, L. J., & Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology*, *16*, 973–980. DOI: 10.1002/acp.920

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. DOI: 10.1186/1471-2105-12-77

Ryan, R. S., & Schooler, J. W. (1998). Whom do words hurt? Individual differences in susceptibility to verbal overshadowing. *Applied Cognitive Psychology*, *12*, 105–125. DOI: 10.1002/(SICI)1099-0720(199812)12:73.0.CO;2-V

Sauerland, M., Holub, F. E., & Sporer, S. L. (2008). Person descriptions and person identifications: Verbal overshadowing or recognition criterion shift? *European Journal of Cognitive Psychology*, *20*, 497–528. DOI: 10.1080/09541440701728417

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*, 989–997. DOI: 10.1002/acp.930

Schooler, J. W., & Engstler-Schooler, T. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71. DOI: 10.1016/0010-0285(90)90003-M

Schooler, J. W., Fiore, S. M., & Brandimonte, M. A. (1997). At a loss from words: Verbal overshadowing of perceptual memories. In D. L. Medin (Ed.), *The psychology of learning and motivation* (pp. 293–334). San Diego, CA: Academic Press.

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement [Booklet]*. Washington, DC: United States Department of Justice, Office of Justice Programs, National Institute of Justice.

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior*, *22*, 217–237. DOI: 10.1023/A:1025746220886

Valentine, T., & Heaton, P. (1999). An evaluation of the fairness of police line-ups and video identifications. *Applied Cognitive Psychology*, *13*, 59–72 [Special issue]. DOI: 10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-Y

Wells, G. L., & Turtle, J. W. (1988). What is the best way to encode faces? In M. M. Gruneberg, P. E. Morris & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues: Memory in everyday life* (Vol. *1*, pp. 163–168). New York: Wiley.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125.

Westerman, D. L., & Larsen, J. D. (1997). Verbal-overshadowing effect: Evidence for a general shift in processing. *The American Journal of Psychology*, *110*, 417–428. DOI: 10.2307/1423566

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*, 275–278. DOI: 10.1177/1745691612442906

Yu, C. J., & Geiselman, R. E. (1993). Effects of constructing identi-kit composites on photospread identification performance. *Criminal Justice and Behavior*, *20*, 280–292. DOI: 10.1177/0093854893020003005