The verbal facilitation effect: re-reading person descriptions as a system variable to improve identification performance

Siegfried L. Sporer, Kristina S. Kaminski, Maike C. Davids and Dawn McQuiston

ABSTRACT
When witnesses report a crime, police usually ask for a description of the perpetrator. Several studies suggested that verbalising faces leads to a detriment in identification performance (verbal overshadowing effect [VOE]) but the effect has been difficult to replicate. Here, we sought to reverse the VOE by inducing context reinstatement as a system variable through re-reading one’s own description before an identification task. Participants (N = 208) watched a video film and were then dismissed (control group), only described the perpetrator, or described and later re-read their own descriptions before identification in either target-present or target-absent lineups after a 2-day or a 5-week delay. Identification accuracy was significantly higher after re-reading (85.0%) than in the no description control group (62.5%) irrespective of target presence. Data were internally replicated using a second target and corroborated by several small meta-analyses. Identification accuracy was related to description quality. Moreover, there was a tendency towards a verbal facilitation effect (VFE) rather than a VOE. Receiver operating characteristic (ROC) curve analyses confirm that our findings are not due to a shift in response bias but truly reflect improvement of recognition performance. Differences in the ecological validity of study paradigms are discussed.

When a crime is reported to the police, a common practice is to ask the victim or witnesses for a description of the perpetrator. Most descriptions contain references to some aspect of the physical appearance of the perpetrator, which are usually not very distinctive, as well as vague estimates of age, height and weight. Among these attributes, the face is the most useful part of the body when trying to identify a person. However, our vocabulary for describing physical attributes of a face is rather limited (cf. Sporer, 1989). Moreover, research has shown that faces are best encoded configurally or holistically (Ellis, 1984; Tanaka & Farah, 2003), whereas giving a description requires accessing the memory of the face by its individual features (e.g., Schooler & Engstler-Schooler, 1990; see Meissner, Sporer, & Susa, 2008). Thus, on the one hand, obtaining a useful description from a witness can be quite difficult (Fahsing, Ask, & Granhag, 2004; Lindsay, Martin, & Webber, 1994; Shepherd, Davies, & Ellis, 1981), while on the other hand being indispensable for furthering the criminal investigation (e.g., Loftus, 1979; Wells & Olson, 2003). In most cases, person descriptions precede a later person identification task. Moreover, the US Supreme Court recommended considering the accuracy of person descriptions when evaluating the reliability of an identification decision (Neil v. Biggers, 1972).

Due to the practical importance of person identifications, research has focussed on the relationships between quantitative and qualitative description measures and identification performance by investigating the post-dictive value of person descriptions to discriminate between correct and incorrect identification decisions (see the meta-analysis by Meissner, Sporer, & Susa, 2008). Other researchers have focused on the impairing effects that giving a description can have on a future identification decision (Schooler & Engstler-Schooler, 1990; see also Meissner & Brigham, 2001).

In contrast, in the present study special emphasis was placed on the potentially beneficial effect of giving a person description on lineup identification accuracy (verbal facilitation), while ensuring the study’s ecological validity. Furthermore, we employ person descriptions as a system variable to improve identification performance (Wells, 1978). Specifically, descriptions were treated and used as self-generated retrieval cues by allowing witnesses to re-read their own descriptions prior to the identification task, thus reinstating the previous retrieval context.
**Description effects: verbal overshadowing and verbal facilitation**

A description given by a witness can be helpful in finding a suspect. But what if the very act of describing the perpetrator impaired a witness’s ability to later identify him or her? Schooler and Engstler-Schooler (1990) found that exactly this might be the case. In their Experiment 1, participants who had been encouraged to give a detailed description of the perpetrator’s face for 5 minutes performed significantly worse (decrease of 25%) in an immediately following target-present identification task than participants in the control condition who did not give a description, an effect the authors termed verbal overshadowing effect (VOE). Recently, Algona et al. (2014) conducted a registered replication report of Schooler and Engstler-Schooler’s (1990) original experiment, including 22 studies. The results supported a robust and consistent VOE with an average difference of 16% between a verbal description condition and a control condition.

However, note that encouraging participants to describe a face for 5 minutes may place undue emphasis on verbalisation, thus provoking potentially misleading descriptors. In a recent study from our laboratory (N = 197) the description of faces, body and clothing lasted on average only $\text{Mdn} = 34$ seconds (interquartile range = 26 seconds). In an archival analysis of person descriptions in criminal cases, only 2.88 face descriptors were mentioned (Sporer, 1996). Hence, we wonder why participant witnesses are encouraged to describe a face for 5 minutes?

Since Schooler and Engstler-Schooler’s (1990) original work, quite a few studies have been conducted on the VOE, some replicating (e.g., Fallshore & Schooler, 1995; MacLin, 2002; Smith & Flowe, 2014; Sporer, 1989), others failing to replicate or even demonstrating a reversal of the effect (e.g., Brown & Lloyd-Jones, 2005, 2006; Chance & Goldstein, 1976; Itoh, 2005; Kitagami, Sato, & Yoshikawa, 2002; McKelvie, 1976; Meissner, Brigham, & Kelley, 2001; Read, 1979; Sauerland, Holub, & Sporer, 2008; Yu & Geiselman, 1993). A major restriction of past studies showing a VOE is that in most of the experiments no or only very short delays of a few minutes were inserted between the description and the identification task. In contrast, in real-world cases, a description–identification interval of several days, or even weeks or months is more likely to occur (cf. the median delay of 13–14 days in a recent field study by Wells, Steblay, & Dysart, 2015). If, however, a delay was inserted between the description and the recognition task, the negative effects of verbalisation disappeared in most studies (e.g., Yu & Geiselman, 1993; see the meta-analysis by Meissner & Brigham, 2001; an exception is Schooler & Engstler-Schooler, 1990, Experiment 5). In a direct test of this argument by Finger and Pezdek (1999), the VOE disappeared when an interval of 24 minutes or 1 hour, respectively, was inserted between description and recognition task. Similarly, in Algona et al.’s (2014) replication report, the VOE was much smaller (4% across 31 studies) when an interval of 20 minutes was used.

These findings were supported in the meta-analysis by Meissner and Brigham (2001). VOE occurred in studies in which the identification task immediately ($Z_r = −.16$), or with a short delay ($Z_r = −.13$) followed the verbalisation task. Differences were marginally significant in the opposite direction in studies employing a delay of more than 30 minutes ($Z_r = .07$), suggesting a verbal facilitation effect (VFE).

Facilitative effects of verbalisation on recognition performance can be explained by Paivio’s (1971) dual process theory of encoding and Craik and Lockhart’s (1972) levels-of-processing account. Both accounts share the assumption that multimodal (verbal and visual) and thus deeper encoding (e.g., by adding self-generated semantic associations while describing the face) should result in a retrieval benefit for the encoded stimulus. As memory strength decreases with time (Ebbinghaus, 1913; Rubin & Wenzel, 1996), we assume that an early first recall, that is, describing the face soon after the witnessed event, leads to a deeper level of processing and elaboration, and thus, consolidates the recalled information into memory and reduces the amount of forgetting (e.g., Ebbesen & Rienick, 1998; see also Hope, Gabbert, & Fisher, 2011). Moreover, in terms of an associative network perspective, an early recall increases the activation level of the recalled items and strengthens the associations between them (Anderson, 1983). Thus, new retrieval routes are produced and related concepts are activated that later can serve as additional retrieval cues. Consequently, later recall attempts are facilitated.

These assumptions are in line with learning studies investigating the beneficial effect of an early memory test prior to the final memory task, called the testing effect. A retrieval-induced facilitation of material related to the tested material is observed when early testing activities are included compared to no-testing control groups (for a review, see McDermott, Arnold, & Nelson, 2014; Karpicke & Grimaldi, 2012; Roediger & Karpicke, 2006), especially with longer delays (e.g., Butler & Roediger, 2007; Chan, 2009). Extrapolating from this literature, one would also expect a (stronger) VFE after longer description–identification delays.

Indeed, several studies have shown a beneficial effect of verbalisation on later face recognition. For example, Sporer (1988) exposed participants to pairs of faces for 10 seconds and told them afterwards, which of the two they were to commit to memory. In one group, participants were telephoned and read their own descriptions to visually rehearse these faces. At a final recognition test, the verbally described and rehearsed faces showed a 12.5% advantage compared to the non-rehearsed faces. It appears that the description fostered the original memory trace. However, most of these studies used old-new recognition paradigms with multiple faces to be described and remembered (e.g., Brown & Lloyd-Jones, 2005, 2006; Brown, Gehrke, &
Lloyd-Jones, 2010; Wickham & Lander, 2008). But most of them did not include realistic post-description delays or used target-present lineups only (e.g., LaPaglia & Chan, 2012). Thus, the main goals of the present study were to investigate description effects using an eyewitness identification paradigm with (1) realistic description instructions, (2) longer description–identification delays and (3) both target-present and target-absent lineups.

**Context reinstatement**

From an associative network perspective (Anderson, 1983), a to-be-remembered stimulus is never encoded into memory alone. Accordingly, a variety of environmental, emotional and other contextual information of the episode in which the stimulus was encountered is encoded in an associative network into which the to-be-remembered stimulus is embedded (Bower, 1981). If a later memory search fails to activate the direct path to the stimulus node, alternative routes can be primed by using contextual cues, making it more likely for the stimulus node to be activated and the required information to be recalled. This effect was termed context reinstatement effect.

In an eyewitness identification study by Cutler, Penrod, and Martens (1987), a context reinstatement interview was used, consisting of Geiselman, Fisher, MacKinnon, and Holland’s (1985) “mnemonic instructions”, pictures of the location and the victim of the incident, and a review of the original description. They found that when the perpetrator was disguised, the context reinstatement interview significantly improved identification performance (51% vs. 29%, $d = .49$, odds ratio $[\text{OR}] = 2.43^3$), whereas it had no significant effect if the perpetrator was non-disguised (47% vs. 57%, $d = -.22$, $\text{OR} = .67$). Beneficial effects of context reinstatement procedures on identification accuracy were observed in a field experiment by Kraffa and Penrod (1985) and a staged event study by Malpass and Devine (1981) when target-present lineups were used (both with $\text{ORs} > 2.00$).

In an early meta-analysis of both facial recognition and lineup identification studies, Shapiro and Penrod (1986) reported a large beneficial effect of context reinstatement on hits ($d = 1.91, k = 23$), but also a smaller increase in false alarms ($d = -.44, k = 18$). Recently, Wong and Read (2011) similarly reported a significant positive effect of context reinstatement on the hit rate in target-present lineups ($\text{OR} = 3.12$), but a non-significant effect on the false alarm rate when the target was absent ($\text{OR} = 1.84$). Consequently, there is a need to develop a method of context reinstatement that will increase hit rates without increasing false identifications.

**Context reinstatement by re-reading one’s own descriptions**

Based on the expected facilitating effect of person descriptions on identification accuracy, the question arises whether one could even further increase this positive effect. Hence, we propose that person descriptions be used as a form of context reinstatement as a simple system variable to further improve identification accuracy.

Cutler, Penrod, O’Rourke, and Martens (1986) attempted to unconfound different context reinstatement procedures and observed that re-reading one’s own description of the target and the event was the only context variable yielding significant effects on identification accuracy. However, beneficial effects of re-reading were present only under certain circumstances, viz., under less optimal retrieval conditions. When the target was absent in the lineup, re-reading had a positive effect (control: 60% vs. re-reading: 74% correct rejections, $d = .30$, $\text{OR} = 1.72$), whereas with target-present lineups it had a negative effect (control: 68% vs. re-reading: 50% hits, $d = -.39$, $\text{OR} = .49$). Moreover, re-reading had positive effects when the perpetrator was disguised and absent from the lineup (control: 50% vs. re-reading: 70% correct rejections, $d = .45$, $\text{OR} = 2.26$).

Sporer (2007) also explored possible effects of re-reading descriptions using a relatively shorter exposure time of the target and a retention interval of 1 week. There was a tendency for participants who re-read their descriptions (52% correct identification decisions) to perform better at the identification task than participants in the no-reread condition (36%), $\text{OR} = 1.90$. However, this effect did not reach significance, due to the small sample size ($N = 54$).

In sum, re-reading one’s own description does not only not impair identification performance, but may actually activate an associative memory network for the target face, resulting in an increase in identification accuracy, especially after long delays and with target-absent lineups. Re-reading descriptions may function as self-generated retrieval cues, which have been shown to induce even higher memory performance compared to other-generated cues (analogous to the studies with word lists: e.g., Mäntylä, 1986). Thus, the present study aimed for a replication of the re-reading effect with new stimulus material to further test its effectiveness, while taking extensive care to ensure ecological validity.

**Do quantity and quality of person descriptions matter?**

The benefit of re-reading is likely to depend on the quantity and quality of a witness’s description. Re-reading should be helpful to the extent that the description includes many (correct) details, which act as retrieval cues to activate the original memory for the target face, thus enabling better identification. Hence, the question is: Does context reinstatement by means of re-reading depend on a “good” person description containing many correct details?

Sporer (1996) identified five aspects that can be related to identification accuracy: the length of the description (i.e., the number of words), the number of details reported, the
accuracy, the internal consistency between different descriptions by the same witness and the general quality of the statement. To judge the “goodness” of a description, both the total number of details reported and the proportion of accurate and inaccurate details have to be considered. Relationships between different aspects of a description and identification are generally weak but stronger if person descriptions are measured with methodological rigor (cf. the meta-analysis by Meissner et al., 2008). Sometimes description properties are not related to identification accuracy but to choosing rates with participants who were allowed to re-read their descriptions prior to the identification task (Sauerland et al., 2008).

This supports the idea that, within re-readers, increased lineup rejections might be due to the perceived inaccuracy of their descriptions making participants more sceptical of their own memory and thus, more reluctant to choose someone from the lineup (cf. the criterion shift account of the VOE: Clare & Lewandowsky, 2004; Sauerland et al., 2008).

In the present study, relationships between identification accuracy and description accuracy as well as the number of details were examined. Because in actual criminal cases, there is no way of assessing the actual accuracy of a description—the true identity of the perpetrator is unknown—different aspects of the perceived description quality (cf. Valentine, Pickering, & Darling, 2003) were additionally measured, including ratings of a description’s precision, specificity and informativeness. Especially for re-readers we expected positive relationships between these ratings and identification accuracy.

The present study

The main goal of the present study was to investigate the effects of verbalisation and re-reading one’s own description on subsequent identification accuracy. We were also interested in examining possible associations with description quality.

Using three groups, a no description control group, a description-only group, and a description plus re-reading group, orthogonally crossed with both target-present and target-absent lineups, allowed us to test rival predictions from the verbal overshadowing theories and the context reinstatement literature. We inserted two ecologically valid delays of 2 days and 5 weeks between the exposure to the target and the identification task, which we predicted would result in a positive effect of verbalisation on recognition performance. Based on accounts of retrieval-based learning (cf. Karpicke & Grimaldi, 2012), we expected greater identification accuracy for participants who gave a target description compared to those in a control group who did not (verbal facilitation hypothesis).

Furthermore, we expected that re-reading one’s description prior to the identification task would serve as a self-generated retrieval cue and, based on an associative memory network model, a mental reinstatement of the encoding context. Re-reading should increase the probability of a correct identification decision compared to a description-only and a no description control group (context reinstatement hypothesis).

To further substantiate this assumption, we investigated whether mock witnesses who had not seen the stimulus film but were only given a person description from a yoked witness-participant would be equally able to make a correct identification decision as participant witnesses who had re-read their own descriptions. Here we sought to rule out the alternative explanation that not the activation of an associative memory network by self-generated retrieval cues was responsible for the expected improvement in identification accuracy but the simple use of anyone’s person descriptions (i.e., other-generated retrieval cues).

Method

Design

To assess the effect of person descriptions and context reinstatement on the accuracy of a subsequent lineup identification decision, two experiments were conducted (see Table 1). In Experiment 1, a 3 × 2 factorial between-participants design was used. Participants were randomly assigned to one of the three groups: a control group in which participants gave no description, a description-only group in which participants provided a description of the perpetrator, and a description re-reading group in which participants provided a description of the perpetrator and were allowed to re-read prior to the identification task that took place 2 days later. The presence versus absence of the target face in the lineup was orthogonally varied. Experiment 2 was identical, however, there was no description-only group and the post-description delay was 5 weeks. Preliminary analyses revealed that the effects reported did not interact with Experiment (1 vs. 2), so we combined the data from both experiments to increase the statistical power.

Participants

Across both experiments, 208 students participated as a course requirement. Ninety-five students were tested at the Arizona State University (32 males and 59 females, 71 females, 36 males), 113 participants were tested at the University of South Carolina (38 males and 75 females).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Experiment 1 Two days delay</th>
<th>Experiment 2 Five weeks delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>No description CG</td>
<td>80 TA 24</td>
<td>80 TA 24</td>
</tr>
<tr>
<td>Description-only</td>
<td>48 24 TA 24</td>
<td>48 16 TA 16</td>
</tr>
<tr>
<td>Re-reading</td>
<td>80 24 TA 24</td>
<td>80 16 TA 16</td>
</tr>
<tr>
<td>Total</td>
<td>208 72 TA 72</td>
<td>208 32 TA 32</td>
</tr>
</tbody>
</table>

Notes: CG = control group; TA = target-absent; TP = target-present. Due to the study’s primary focus on the effects of re-reading, there was no description-only group in Experiment 2.
age 17–52, $Mdn = 20.0$), and 117 at the Justus-Liebig University Giessen in Germany (32 males and 85 females, age 19–53, $Mdn = 22.0$). The participants were mainly Caucasian. Participants of other ethnic groups (Mexican-American and Native American, Asian, African) originated from the American sample and were equally distributed across the different conditions. Participants were tested in groups of one to five persons and individually seated in front of personal computers. Preliminary analyses revealed that there were no differences in results across countries. Therefore, location of the study will not be considered as a factor.

**Materials**

**Stimulus film**

The stimulus was a high-quality colour and sound video film showing a theft of a wallet from a young man’s backpack. It was filmed at a sidewalk in a quiet residential area and had two actors. A young Caucasian male (“victim”), 22 years old, was searching through his backpack when he was approached by another male (“thief”), 25 years old, who asked for directions. After the victim had finished giving his directions, his cell phone rang and he walked a few steps away to take the call, turning his back to the thief and leaving his backpack lying on the floor. The thief then quickly bent down and took the victim’s wallet out of the backpack. He thanked the victim for the directions and walked off. The victim was still speaking on the phone, saying: “Yes, I’ve got her number, it’s in my wallet, hold on, I’ll get it for you.” He then went back to his backpack and searched for his wallet. When he could not find it, he took several things (water bottle, book, sweater, sheets of paper, keys) out of his backpack. When his wallet still did not show up, he angrily got up and told the person on the phone: “I’ll get it for you.”

**Procedure**

After arriving at the laboratory, participants were seated in front of a 35-inch TV screen. Participants were instructed to watch the film attentively because they were going to be asked several questions about it afterwards. Following the film, control group participants ($n = 80$) were allowed to leave.

**Description task**

Shortly after seeing the film, participants of the two experimental groups ($n = 128$) were seated in front of the 15-inch screen of a Macintosh computer. The program SuperLab 1.75 (www.cedrus.com) was used for all instructions, lineup tests and data collection. At first, participants were instructed to imagine having to give a statement about the incident seen in the film to the police. They were asked to describe the incident in as much detail as possible on an answer sheet. Following this, they were asked some non-leading specific questions about the event (approximately 5–10 min). After that, they first gave a free description of the perpetrator, followed by 12 specific, non-leading questions about his appearance (see Appendix). Participants were instructed to describe the perpetrator as precisely as possible on an answer sheet, so that another person could find him in a crowd. Importantly, they had the opportunity to give “don’t know” answers, if they were not able to remember any of the specific features asked for. In Experiment 1, participants were also asked to describe the victim with equivalent instructions and questions.

**Re-reading manipulation and identification tasks**

Two days ($n = 144$) or 5 weeks ($n = 64$) later all participants, including the control group, returned to the laboratory. All participants were presented with the lineup task on a 15-inch computer screen. Prior to the lineup task, only participants in the re-reading group were allowed to read again their own free description of the thief they had given earlier. All participants were presented with unbiased lineup instructions, which stressed the possibility that the thief may or may not be present in the lineup. Following this, they either saw a TP or TA lineup ($n = 104$ in each condition). Afterwards, all participants had to give a confidence rating about their lineup decision on an 11-point scale (0–100%). Next, the same procedure was repeated for the identification of the victim (in Experiment 1 only). At the end, participants were thanked, asked not to talk about the experiment with future participants and released. All procedures were in accordance with departmental (based on APA) ethical guidelines.

**Rating and coding of person descriptions**

The free reports as well as the specific questions regarding the perpetrator’s physical appearance were analysed to obtain separate scores for description quantity (total...
number of details), the number of correct details, the number of false details and description accuracy (= number of correct details/number of correct details plus number of false details)). A comprehensive coding scheme was prepared in a pilot study, which was used by two independent coders to code the descriptions.

**Preparation of a coding scheme**

First, three persons were instructed to watch the stimulus film carefully and to create as many feature categories considered necessary to capture all possible aspects of the perpetrator’s physical appearance that could have been described by any participant of the main study. This resulted in a total of 130 features of the perpetrator’s face, hair, body, clothes and accessories.

Then, in a pilot study $N = 20$ participants individually watched the stimulus film and were asked to describe the perpetrator based on those preset feature categories, as precisely as possible (e.g., “Which eye-color does the perpetrator have?” or “Describe the perpetrator’s skin texture.”). They were allowed to watch the film as often as necessary and/or to stop it anytime to find all the information needed to answer the questions. For every of the 130 features, the most frequently stated answer (the *mode*) was defined as correct. In most cases, there were clear modal answers (e.g., 18 out of 20 persons described the perpetrator’s eyes as “brown”, thus brown eyes were adopted as correct answer in the coding scheme). If there was no clear majority in the participants’ answers or if there were any ambiguities in the formulations, two additional coders had to agree on the correct answer. Age was coded correct if it matched the perpetrator’s true age (25 years ± 1 year). For perpetrator’s height and weight, answers within a specific range were defined that should be coded as correct (184.5–190.5 cm and 77.5–82.0 kg, respectively). Gender was not coded.

The final coding scheme resulted in a total of 132 variables (one variable each for age, height and weight, as well as 51 variables describing the perpetrator’s face, 8 variables describing his hair, 18 his body, 26 his clothes and 26 describing further accessories).

**Ratings of description quality**

First, all descriptions in the free report were rated on three different dimensions regarding their content quality by two independent coders. Ratings referred to a description’s precision, specificity and informativeness using 7-point Likert scales (i.e., 1 = not at all precise/specific/informative; 7 = very precise/specific/informative). Precision was defined as a measure for the elaborateness and clearness of the description. If a description was rated as highly precise, lots of different features were explicitly stated (e.g., facial form, nose, mouth, skin, eyes, ears, hair structure, hair colour and clothes). In contrast, descriptions containing just a few vaguely described features were to be rated low in precision (e.g., hair colour and length). Specificity explicitly referred to the degree of differentiation the features were described with (e.g., “Hair was short, 5–7 cm, brown and curly” vs. “brown hair”). Informativeness referred to the description’s ability to differentiate the perpetrator from other persons (i.e., to find him in a crowd). Thus, in a highly informative description, the perpetrator’s unique or distinctive features were emphasised (e.g., “He had a small tattoo on his right arm.”).

**Coding of descriptions for number and accuracy of details**

Descriptions given in free report and in specific questions were coded separately. The coding scheme as well as the descriptions of $N = 101$ participants from the main study were imported into Maxqda2 (www.maxqda.de), a software program for qualitative data and text analysis. Two independent coders who were familiar with the complex coding procedure, coded every descriptor as “correct”, “incorrect”, “confabulated” or “subjective”. If a feature was mentioned that the perpetrator did not have (e.g., he wore glasses, although he did not wear any), the item was treated as a confabulation. Items coded as subjective were descriptions idiosyncratic for a participant (e.g., “He looked like my brother”; “He was handsome”). For details not included in the coding scheme, a “rest” category was used.

After the coding process, incorrect and confabulated items were merged into a category of “false” descriptors due to low frequencies. Categories of subjective details and remaining features were excluded from further analyses for the same reason (in the free reports, <50% and <25% of participants mentioned one or more features that were coded in these categories).

**Subsequent data collection**

To eliminate the mere use of descriptions as an alternative explanation of the proposed re-reading effect, an additional $N = 128$ participants (76 males and 52 females, age 16–79, $Mdn = 26.0$), who did not see the stimulus film, were instructed to identify the perpetrator solely based on the 128 free person descriptions given by the participants in the two main experiments. First, every participant had to read one yoked participant’s free target description, and was then immediately shown the same lineup the corresponding participant of the main experiments had seen. Participants received the same unbiased lineup instructions used in the main study, that is, they were informed about the possibility to reject the lineup.

**Results**

**Preliminary analyses: description use as alternative explanation for the expected re-reading effect?**

To rule out the possibility that simply reading anyone’s description might be responsible for correct identification decisions, we compared identification accuracy of the
two re-reading groups from both experiments and mock witnesses who had not seen the film but were asked to select the target based on a yoked description of an experimental participant. Identification accuracies were clearly higher for re-readers (85.0%) than mock witnesses (37.5%), LR $\chi^2(1, N = 160) = 40.15$, $p < .001$ and $OR = 9.44$. Similarly, participants in the description-only group of Experiment 1 (79.2%) made significantly more correct identification decisions than mock witnesses given these descriptions (33.3%), LR $\chi^2(1, N = 96) = 21.35$, $p < .001$ and $OR = 7.60$.

Overview of further analyses

Preliminary analyses revealed that the results of Experiments 1 and 2 were completely parallel, with the exception of the influence of delay. Therefore, we first present the joint results of Experiment 1 and 2 regarding the potential benefit of re-reading on the accuracy of identification decisions. Subsequently, we analyse the effect of prior describing a perpetrator on subsequent identifications (i.e., no description control group [CG] vs. the two description groups combined) to differentiate between verbal overshadowing versus verbal facilitation effects.  

In light of the ongoing controversy regarding the most appropriate analysis of identification data, we first report diagnosticity ratios (DR) as well as the signal-detection theory based analyses of ROC curves (cf. Gronlund, Mickes, Wixted, & Clark, 2015; Gronlund, Wixted, & Mickes, 2014; Mickes, Flowe, & Wixted, 2012; Mickes, Moreland, Clark, & Wixted, 2014; Wells, Smalarz, & Smith, 2015; Wells, Yang, & Smalarz, 2015; Wixted & Mickes, 2012, 2014). We are aware that conducting repeated comparisons potentially inflate the type-1 error rate. However, the main purpose here was to demonstrate that similar conclusions can be drawn no matter which of these analyses are conducted.

Further, data were analysed via four hierarchical log-linear frequency analyses (SPSS Hiloiglinear, hierarchical backward elimination method), which are considered appropriate when both independent and dependent variables are categorical in nature (Howell, 2013; Tabachnik & Fidell, 2007; cf. also Meissner et al., 2001).

To describe the size of effects for categorical variables, we report OR, that is, the ratio of the odds for a given outcome (e.g., the odds for a correct identification decision) in one group divided by the odds for the same outcome in another group (cf. Fleiss, 1994; Lipsey & Wilson, 2001). For example, in the re-reading group, the odds of a correct lineup decision are the proportion of correct decisions (e.g., .85) divided by the proportion of incorrect decisions (e.g., .15), which equals 5.7. Thus, in this condition, the odds for correct decisions are almost six times higher than those for incorrect decisions. In contrast, in the no description CG the odds for a correct decision were .63/.37 = 1.7. The OR is the ratio between these two odds, which is 5.7/1.7 = 3.4 in this case. Hence, the odds for a correct lineup decision in the re-reading group are 3.4 times higher than in the no description CG. An OR of 1 means the odds of correct lineup decisions to be the same for both groups.

At the end, point-biserial correlations between the different description measures and identification accuracy as a function of delay and experimental condition (description-only vs. re-reading group) are reported.

Diagnosticity ratios versus receiver operating characteristic analyses

Diagnosticity ratios (= correct identifications in target-present lineups/false identifications in target-absent lineups/lineup size), signal-detection theory performance $d'$ and response criterion $C$ (cf. Macmillan & Creelman, 2005) for the different description groups are displayed in Table 2. The indices $d'$ and $C$ were calculated based on hit and false alarm rates, with foil identifications in TP lineups excluded from the analysis (cf. Meissner, Tredoux, Parker, & MacLin, 2005). False alarm rates were calculated in the same way as for diagnosticity ratios, that is, false identifications in TA lineups were divided by nominal lineup size (see “Estimated FAs” in Table 2).

Moreover, ROC curves for each description condition were constructed based on the current approach illustrated by Gronlund et al. (2014) using participants’ identification response data and confidence ratings. Therefore, the cumulative hit and false alarm rates (HR and FAR) were plotted at each confidence level, ranging from conservative responding at the one end (i.e., identifications with high confidence levels only) to a more liberal response criterion at the other end. Again, false alarm rates were corrected by the nominal size of the lineup, and foil identifications in TP lineups were excluded from the analysis. Then, partial area under the curve (pAUC) analyses were conducted to compare identification discriminability across the different description conditions. The procedure yielding the higher pAUC is considered diagnostically superior, that is, yielding a higher ability to distinguish between innocent and guilty suspects in a lineup. Partial AUC analyses are appropriate here, because the data do not include the full range of HR and FAR from 0 to 1. For each comparison, we selected the maximum FAR as the cutoff point. pAUCs were computed and compared using the data analysis package pROC for R (Robin et al., 2011), applying the bootstrapping method (with the number of bootstraps set to 10 000). For comparing the two pAUCs, the following formula was used: $D = (AUC1 − AUC2)/s$, 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 (Robin et al., 2011).

Parallel to our further analyses described below, we first compared the identification performance of the no description control group and the re-reading groups in both experiments. The pAUC of the no description control group (pAUC = .012, CI$._{95\%}$ = .006–.022) was
significantly smaller than the pAUC of the re-reading groups (pAUC = .031, CI\textsubscript{95\%} = .021–.041), \(D = -2.75\), \(p = .006\). The ROC curves are displayed in Figure 1, whereby FARs were divided by the total number of persons in the lineup (i.e., six).

Next, possible effects of verbalisation on identification discriminability (VOE or VFE) were tested. The difference between the pAUCs of both description conditions joined (pAUC = .024, CI\textsubscript{95\%} = .015–.037) and the no description control group (pAUC = .012, CI\textsubscript{95\%} = .006–.022) was only marginally significant, \(D = -1.85\), \(p = .064\).

In Experiment 1, comparing the pAUCs of the description-only group (pAUC = .012, CI\textsubscript{95\%} = .007–.024) with the no description control group (pAUC = .010, CI\textsubscript{95\%} = .005–.023), showed no significant difference, \(D = -.26\), \(p = .798\).

### Effects of re-reading on identification accuracy

**Comparing the re-reading with the no description control group**

First, we focus on the conditions present in both experiments (\(N = 160\)), thus excluding the description-only condition in Experiment 1. Table 2 and Figure 2 give an overview of the accuracy of identification decisions and other outcomes in the different conditions in both experiments.

Data were analysed via a four-way hierarchical log-linear frequency analysis including re-reading condition (no description CG vs. re-reading group), delay (2 days vs. 5 weeks), target-presence (TA vs. TP) and accuracy of identification decisions (coded as a binary variable: \(0 = \text{incorrect} \) [false rejection or filler identification in TP lineups; any

#### Table 2. Identification outcomes (% with frequencies in parentheses), diagnosticity ratios, signal-detection performance \(d'\) and response bias \(C\) for description conditions in Experiment 1 (2 days delay) and Experiment 2 (5 weeks delay).

<table>
<thead>
<tr>
<th>Condition</th>
<th>(n)</th>
<th>Correct decisions</th>
<th>Hits</th>
<th>FAs</th>
<th>Estimated FAs</th>
<th>DR</th>
<th>(d')</th>
<th>(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No description CG</td>
<td>48</td>
<td>77.1 (37)</td>
<td>75.0 (18)</td>
<td>20.8 (5)</td>
<td>3.5</td>
<td>21.6</td>
<td>2.49</td>
<td>.57</td>
</tr>
<tr>
<td>Description only</td>
<td>48</td>
<td>79.2 (38)</td>
<td>70.8 (17)</td>
<td>12.5 (3)</td>
<td>2.1</td>
<td>34.0</td>
<td>2.58</td>
<td>.74</td>
</tr>
<tr>
<td>Re-reading</td>
<td>48</td>
<td>91.7 (44)</td>
<td>87.5 (21)</td>
<td>4.2 (1)</td>
<td>.7</td>
<td>125.0</td>
<td>3.61</td>
<td>.65</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No description CG</td>
<td>32</td>
<td>40.6 (13)</td>
<td>25.0 (4)</td>
<td>43.8 (7)</td>
<td>7.3</td>
<td>3.4</td>
<td>.78</td>
<td>1.06</td>
</tr>
<tr>
<td>Description only</td>
<td>32</td>
<td>75.0 (24)</td>
<td>56.2 (9)</td>
<td>6.2 (1)</td>
<td>1.0</td>
<td>54.4</td>
<td>2.48</td>
<td>1.09</td>
</tr>
<tr>
<td>Re-reading</td>
<td>32</td>
<td>62.5 (50)</td>
<td>55.0 (22)</td>
<td>30.0 (12)</td>
<td>5.0</td>
<td>11.0</td>
<td>1.77</td>
<td>.76</td>
</tr>
<tr>
<td><strong>Experiments 1 and 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No description CG</td>
<td>80</td>
<td>65.0 (50)</td>
<td>55.0 (22)</td>
<td>30.0 (12)</td>
<td>5.0</td>
<td>11.0</td>
<td>1.77</td>
<td>.76</td>
</tr>
<tr>
<td>Description only</td>
<td>48</td>
<td>79.2 (38)</td>
<td>70.8 (17)</td>
<td>12.5 (3)</td>
<td>2.1</td>
<td>34.0</td>
<td>2.58</td>
<td>.74</td>
</tr>
<tr>
<td>Re-reading</td>
<td>80</td>
<td>85.0 (68)</td>
<td>75.0 (30)</td>
<td>5.0 (2)</td>
<td>.8</td>
<td>90.0</td>
<td>3.08</td>
<td>.87</td>
</tr>
</tbody>
</table>

Notes: Hits = correct identifications in TP lineups; FAs = any positive identification in TA lineups; estimated FAs = false identifications in TA lineups/lineup size; DR = diagnosticity ratio.

#### Figure 1.
Confidence-based receiver operating characteristic (ROC) curves for the re-reading and no description control group. Data points reflect cumulative hit and false alarm rates (i.e., false alarms in TA lineups/lineup size) for each confidence level. The rightmost point of the ROC represents the hit and false alarm rate across all confidence levels (i.e., most liberal decision criterion). The remaining points were computed using ever lower cutoff values on the confidence scale (i.e., an increasingly conservative decision criterion).

#### Figure 2.
Means (and 95% CIs) for identification accuracy (hits and correct rejections) in the no description control, the description-only and the re-reading condition after a 2-day versus 5 weeks delay. Note that there was no description-only group after 5 weeks.
positive identification in TA lineups; \( I = \text{correct} \) [hit in TP lineups or correct rejection in TA lineups]).

Stepwise selection by simple deletion of effects produced a model that included only first-order effects and two-way associations. Tests of k-way interactions revealed that none of the three-way or four-way interactions were significant. The final model had a likelihood ratio \( \chi^2(8, N = 160) = 5.93, p = .655 \), indicating a good fit between observed and expected frequencies generated by the model. There were highly significant partial associations between identification accuracy and re-reading, partial \( \chi^2(1, N = 160) = 12.37, p < .001 \), identification accuracy and delay, partial \( \chi^2(1, N = 160) = 15.67, p < .001 \), and identification accuracy and target-presence, partial \( \chi^2(1, N = 160) = 7.62, p = .006 \). Re-reading one’s prior description improved identification accuracy from 62.5% in the no description CG to 85.0% in the re-reading group (OR = 3.40). Thus, the odds of correct identification decisions were 3.4 times higher in the re-reading group than in the no description CG. Performance deteriorated from 84.4% after 2 days to 57.8% after 5 weeks (OR = 3.94). Perhaps surprisingly, participants were better with TA lineups than when tested with TP lineups (82.5% vs. 65.0%, OR = .39).

Although there was no evidence for the existence of three-way interactions, it is worth noting that the benefit of re-reading was at least as strong, if not stronger, in TA lineups (95.0% vs. 70.0% correct rejections; OR = 8.14) as in TP lineups (75.0% vs. 55.0% hits; OR = 2.45). Notably, the effect was comparable in magnitude for the 2-day (91.7% vs. 77.1% correct decisions; OR = 3.27) and 5-week delay (75.0% vs. 40.6% correct decisions; OR = 4.38).

**Comparing the re-reading with the description-only group**

Although re-readers apparently yielded slightly better identification outcomes than participants in the description-only group (see Table 2), none of the differences, neither for TP nor for TA lineups, were statistically significant, all \( p ’ s > .148 \).

**Effects of verbalisation on identification accuracy: verbal facilitation or verbal overshadowing?**

The data presented above demonstrate that re-reading one’s description facilitates performance rather than impairing it. In Experiment 1, a description-only control group was used to test whether this beneficial effect is restricted to re-reading or if giving a description itself is sufficient to increase identification accuracy.

**Comparing both description groups of Experiment 1 and 2 with the control group**

Data were analysed via a four-way hierarchical log-linear frequency analysis including description condition (no description [CG] vs. description [description-only plus re-reading]), delay (2 days vs. 5 weeks), target-presence (TA vs. TP) and identification accuracy (0 = incorrect; 1 = correct). The likelihood ratio for the final model was \( \chi^2(2, N = 208) = 45, p = .098 \), revealing a good model fit. Tests of k-way interactions revealed that all three-way or four-way interactions could be excluded from the model. Besides the already known highly significant associations between identification accuracy and delay as well as target presence (all \( p ’ s < .002 \)), there was a significant two-way effect of Description Condition \( \times \) Identification Accuracy, partial \( \chi^2(1, N = 208) = 8.06, p = .005 \). Combining the description-only group of Experiment 1 (79.2%) and the two re-reading groups (85.0%; \( M \) of both conditions = 82.8%) and comparing them with the no description CG (62.5%) yielded a highly significant VFE (OR = 2.89). Thus, when person descriptions were given, the odds of correct identification decisions were almost three times higher than in the no description CG.

Although the three-way interaction was not significant, separate analyses within TA and TP lineups yielded similar conclusions. In the TA condition, performance was significantly higher in the two description conditions combined (92.2%) than in the control condition (70.0%), LR \( \chi^2(1, N = 104) = 8.67, p = .003, OR = 5.06 \). In the TP condition, the effect failed to reach significance, with a tendency for higher identification accuracies in the description conditions (73.4%) than in the no description CG (55.0%), LR \( \chi^2(1, N = 104) = 3.71, p = .054, OR = 2.26 \).

**Comparing the description-only with the control group**

Only in Experiment 1, it was possible to test the traditional VOE by comparing the description-only group (79.2%) with the no description CG (77.1%), which was far from significant, LR \( \chi^2(1, N = 96) = .06, p = .805, OR = 1.13 \). Separate analyses for TA and TP lineups also revealed neither a VOE nor a VFE effect, both \( p ’ s > .40 \) (correct rejections: OR = 1.84; hits: OR = .81; see the means in Table 2).

**Replication of the effects with another target: victim identification accuracy**

To replicate the observed effects of re-reading and verbalisation on identification accuracy, parallel analyses were conducted for the identification of the victim in Experiment 1. There was a significant effect of re-reading, LR \( \chi^2(1, N = 96) = 6.25, p = .012, OR = 2.87 \), with 70.8% correct identification decisions for re-readers compared to 45.8% for the no description CG. Moreover, participants who gave a victim description (re-reading plus description-only group combined) performed better at the identification task (67.7%) than the no description CG (45.8%), LR \( \chi^2(1, N = 144) = 6.35, p = .012, OR = 2.48 \). There was a marginally significant VFE without re-reading, with 64.6% of the description-only group making a correct lineup decision compared to 45.8% of the control group, LR \( \chi^2(1, N = 96) = 3.43, p = .064, OR = 2.16 \).
Quantity and quality of person descriptions as “postdictors” of identification accuracy

Results are presented separately for free reports and specific questions. As only free descriptions were re-read prior to the identification task, we focused on the relationship between free report description measures and identification accuracy. Features mentioned twice (i.e., in both free report and in specific questions) were not coded separately, and thus, no overall description score was calculated.

Details coded for different feature categories (face, hair, body, clothes and accessories) were summed up for the total number of details reported regarding the whole person.

Inter-coder reliabilities

Different measures of inter-coder reliability were computed separately for free reports and for specific questions (Pearson correlation coefficient \( r \) and intra-class correlation coefficients ICC). All values were highly satisfactory.

Pearson correlations between the two coders for the total number of descriptors in free reports were \( r(99) = .90 \), and \( r(99) = .93 \) for specific questions (both ICCs = .95). Codings for the number of correct details correlated by \( r(99) = .89 \) in free reports, and by \( r(99) = .92 \) in specific questions (both ICCs = .94). Correlations for the number of false details were \( r(99) = .71 \) (ICC = .83), and \( r(99) = .82 \) (ICC = .90), respectively. Accuracy scores correlated by \( r(99) = .67 \) in free reports, and by \( r(99) = .87 \) in specific questions (ICCs = .81 and .91, respectively).

For the ratings, reliabilities were satisfactory as well (precision: \( r(99) = .66 \), ICC = .78; specificity: \( r(99) = .62 \), ICC = .75; informativeness: \( r(99) = .61 \), ICC = .71). For further analyses, the means of the two coders’ ratings were used.

Descriptive statistics of description quantity and quality

In free reports, \( M = 10.68 \) (SD = 4.18) details were mentioned concerning the perpetrator’s physical appearance. Thereof, \( M = 7.51 \) (SD = 3.76) descriptors were coded as correct, and \( M = 3.17 \) (SD = 1.87) descriptors as false. This resulted in an overall accuracy rate of 69.14% (SD = 17.74). In specific questions, participants reported \( M = 14.05 \) (SD = 3.57) details in total, whereof \( M = 8.79 \) (SD = 3.42) were coded as correct and \( M = 5.26 \) (SD = 1.98) were coded as false. Accuracy rate was \( M = 61.76% \) (SD = 14.01). The distribution of details across the different feature categories (face, hair, body, clothes and accessories), separately for free report and specific questions is displayed in Table 3. Specific questions yielded more correct and more false details but lower accuracy than free reports, all \( p \)'s < .001.

Table 3. Means and standard deviations of description quantity, number of correct and false descriptors and description accuracy in free reports and in specific questions (\( N = 101 \)).

<table>
<thead>
<tr>
<th>Category</th>
<th>Free report</th>
<th>Specific questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Total number of details</td>
<td>2.96</td>
<td>1.90</td>
</tr>
<tr>
<td>Face</td>
<td>2.13</td>
<td>.73</td>
</tr>
<tr>
<td>Hair</td>
<td>.96</td>
<td>.85</td>
</tr>
<tr>
<td>Body</td>
<td>4.63</td>
<td>2.79</td>
</tr>
<tr>
<td>Sum</td>
<td>10.68</td>
<td>4.18</td>
</tr>
<tr>
<td>Number of correct details</td>
<td>2.18</td>
<td>1.65</td>
</tr>
<tr>
<td>Face</td>
<td>1.88</td>
<td>.75</td>
</tr>
<tr>
<td>Hair</td>
<td>.28</td>
<td>.49</td>
</tr>
<tr>
<td>Body</td>
<td>3.18</td>
<td>2.66</td>
</tr>
<tr>
<td>Clothes</td>
<td>7.51</td>
<td>3.76</td>
</tr>
<tr>
<td>Number of false details</td>
<td>.78</td>
<td>.87</td>
</tr>
<tr>
<td>Face</td>
<td>.25</td>
<td>.48</td>
</tr>
<tr>
<td>Hair</td>
<td>.68</td>
<td>.77</td>
</tr>
<tr>
<td>Body</td>
<td>1.46</td>
<td>1.24</td>
</tr>
<tr>
<td>Clothes</td>
<td>3.17</td>
<td>1.87</td>
</tr>
</tbody>
</table>

Relationship between identification performance and ratings of description quality

Finally, we examined whether identification performance co-varied with the rated degree of description precision, specificity and informativeness. Additionally, based on high inter-correlations between these variables, we computed a mean composite rating across all four ratings (Cronbach’s alpha = .95). Point-biserial correlations are displayed in Table 5. In the re-reading group, each rating variable was significantly positively correlated with identification accuracy. In contrast, for the description-only group none of these correlations were reliable.

Discussion

The present study aimed to investigate the potentially beneficial effect of describing a target’s physical appearance on a subsequent identification. In particular, we explored the usefulness of re-reading one’s own free target description prior to the identification task as a
new system variable. As a possible explanation for the expected beneficial effect of re-reading, the relationships between identification performance and description quantity, accuracy as well as quality ratings were considered. A primary concern was to ensure ecological validity. Hence, we inserted a retention interval of either 2 days or 5 weeks between observation of the crime and the identification task. Person descriptions were collected according to common police practice, that is, a free report was followed by open-ended questions. By allowing a “Don’t Know” option, we discouraged the self-generation of false details. Both unbiased TA and TP lineups were used to assess the effects on identification accuracy.

**Verbal facilitation versus verbal overshadowing**

Contrary to assumptions in the VOE literature (e.g., Meissner & Brigham, 2001; Meissner et al., 2008; Schooler & Engstler-Schooler, 1990), identification performance across all description groups, with or without re-reading (83%) compared to a no description control group (63%), was superior. The odds of correct identification decisions were almost three times larger when prior descriptions were given compared to the no description group (OR = 2.89).

The effect was especially strong for TA lineups (correct rejections: 93% vs. 70%, OR = 5.06), while there was only a non-significant tendency of facilitation when the target was present (hits: 73% vs. 55%, OR = 2.26). Even when re-readers were not considered in the analysis, thus comparing the description-only group with the no description CG in Experiment 1, results showed that verbalisation per se did not impair identification performance (OR = 1.13). Note that Meissner and Brigham’s (2001) meta-analysis also showed a VFE after the insertion of a post-description delay of more than 30 minutes. Hence, our results support the assumption that realistic delays of more than 2 days or even several weeks, in combination with conservative description instructions, seem to annihilate or even reverse the VOE.

The present results are in line with the assumed memory advantages that are accompanied by early first retrieval attempts (cf. the testing effect; McDermott et al., 2014; Roediger & Karpicke, 2007), especially after longer delays (cf. Butler & Roediger, 2007; Chan, 2009).

In sum, these findings suggest that under ecologically valid conditions, if identification performance is affected by verbalisation at all, it seems to be in a positive, and not, as assumed by the VOE hypothesis, in a negative way. Longer retention intervals of several days or weeks are more representative of real crime situations (see Sporer, 1996; Wells, Steblay, & Dysart, 2015).

**Re-reading as an effective retrieval cue after a long delay**

As predicted, we found a positive effect of context reinstatement by re-reading one’s description prior to the identification task. Identification accuracy was higher for participants that were allowed to re-read their description prior to identification (85%) than for participants who were not allowed to do so (63%), with odds for re-readers to make an accurate identification decision more than three times as high than the odds for non-re-readers (OR = 3.40). The effect was internally replicated for the identification of the victim (OR = 2.87).

Moreover, diagnosticity ratios were much higher in the re-reading group (90.0) compared to the no description control group (11.0). Thus, after re-reading witnesses were 90 times more likely to identify the target than to choose an innocent suspect. Due to the current criticisms concerning diagnosticity ratios (e.g., Wixted & Mickes, 2012), signal detection based ROC analyses were computed, which led exactly to the same conclusions.

The observed beneficial effect of re-reading tended to be stronger for TA (OR = 8.14) than for TP lineups (OR = 2.45). Thus, there was a tendency for re-reading to be more effective to reduce false identifications in TA.

---

**Table 4.** Point-biserial correlations between identification accuracy and description quantity, number of correct and false descriptors and description accuracy for free reports and specific questions.

<table>
<thead>
<tr>
<th>Details</th>
<th>Sample</th>
<th>n</th>
<th>Number</th>
<th>Correct</th>
<th>False</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free report</td>
<td>Total sample</td>
<td>101</td>
<td>.23*</td>
<td>.22*</td>
<td>.06</td>
<td>.06</td>
</tr>
<tr>
<td>Delay</td>
<td>2 days</td>
<td>69</td>
<td>.13</td>
<td>.15</td>
<td>−.03</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>5 weeks</td>
<td>32</td>
<td>.34*</td>
<td>.30*</td>
<td>.15</td>
<td>−.01</td>
</tr>
<tr>
<td>Condition</td>
<td>Description only</td>
<td>33</td>
<td>.13</td>
<td>.13</td>
<td>.04</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>Re-reading</td>
<td>68</td>
<td>.29*</td>
<td>.28*</td>
<td>.09</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td>Specific questions</td>
<td>68</td>
<td>.26*</td>
<td>.24*</td>
<td>.06</td>
<td>.10</td>
</tr>
</tbody>
</table>

Note: Number = number of details; Correct = number of correct details; False = number of false details.

*p < .05.

---

**Table 5.** Point-biserial correlations between quality ratings of perpetrator descriptions in free reports with identification accuracy.

<table>
<thead>
<tr>
<th>Sample</th>
<th>n</th>
<th>Spec</th>
<th>Info</th>
<th>Prec</th>
<th>Comp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sample</td>
<td>101</td>
<td>.14</td>
<td>.16</td>
<td>.17*</td>
<td>.16</td>
</tr>
<tr>
<td>Delay</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two days</td>
<td>69</td>
<td>.07</td>
<td>.04</td>
<td>.08</td>
<td>.07</td>
</tr>
<tr>
<td>Five weeks</td>
<td>32</td>
<td>.17</td>
<td>.31*</td>
<td>.24</td>
<td>.25</td>
</tr>
<tr>
<td>Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description only</td>
<td>33</td>
<td>.00</td>
<td>.02</td>
<td>.08</td>
<td>.03</td>
</tr>
<tr>
<td>Re-reading</td>
<td>68</td>
<td>.25*</td>
<td>.28*</td>
<td>.24*</td>
<td>.27*</td>
</tr>
</tbody>
</table>

Note: Spec = specificity; Info = informativeness; Prec = precision; Comp = mean composite rating.

*p < .10.

*p < .05.
lineups than to increase hits in TP lineups. The positive re-reading effects did not differ as a function of the different delays, a finding future studies should explore – with even longer retention intervals. As re-readers and participants of the description-only group did not differ in their identification performance, a joint effect of giving a person description and re-reading is assumed. However, based on Experiment 1, describing the perpetrator’s face alone is not sufficient to explain the observed improvement in identification accuracy on its own.

Furthermore, the observed re-reading benefit seems not just to be due to the use of an appropriate person description. As the mock witness data reveal, merely reading someone else’s person description does not suffice to select the target in a lineup (nor to correctly reject it). Mock witnesses who did not watch the stimulus film but only read someone else’s person description yielded much lower identification accuracy. Thus, our results are compatible with the view that re-reading one’s own description of the target does induce processes akin to spreading activation in a memory network in which the original target face is embedded (cf. Anderson, 1983). In line with Paivio’s (1971) dual coding hypothesis, we assumed that describing the target should lead to coding features about the target not only visually but also verbally. Presumably, re-reading the description prior to identification would lead to a reinstatement of this verbal code as a self-generated retrieval cue as well as to a re-activation of the associated encoding context resulting in better recognition performance.

To further understand the underlying mechanism that affects person memory and identification, future research should examine whether the observed benefits due to re-reading are restricted to the use of one’s own description as a self-generated retrieval cue compared to being given another person’s (e.g., a co-witness’s) description (other-generated cues; cf. Mäntylä, 1986; Sporer, 1991).

**Robustness of the re-reading effect: a meta-analysis**

Several studies have found some form of re-reading effect (Cutler et al., 1986; Cutler et al., 1987; Sauerland et al., 2008; Sporer, 2007). To demonstrate the robustness of our findings including conceptual replications with different stimulus materials (films, targets, and lineups), we conducted several small meta-analyses, which include the four studies mentioned above as well as the present results. Following the procedures of Lipsey and Wilson (2001), mean weighted effect sizes (ORs) for different lineup outcomes (overall correct identification decisions [TA and TP], hits [TP] and correct rejections [TA]) were calculated, comparing re-reading versus no description control groups (Table 6) and re-reading versus description-only groups (Table 7). All effects for the comparison between a re-reading and a no description CG were reliable, with ORs between 2 and 3. Regarding the comparison between a re-reading and a description-only group, there was only a beneficial effect for correct rejections (OR = 1.79), which seems to be most important regarding preventing mistaken identifications. Moreover, the results support the idea that verbalisation and re-reading make people aware of their poor memory of the target, which in turn may make them more cautious at the identification task (cf. Sauerland et al., 2008; Sporer, 2007).

Hence, re-reading one’s own description may be worth considering as a simple and easy to implement system variable to improve identification accuracy.

**Quantity and quality of person descriptions as possible reasons for the re-reading effect**

To test the assumption that the beneficial effect of re-reading might be associated with the quantity and quality of the re-read person descriptions, relationships between description measures and identification accuracy were investigated.

**Description quantity and accuracy**

Different from the findings in Meissner et al.’s (2008) meta-analysis, in free reports and in specific questions we found moderate positive associations between identification accuracy with the number of details reported (cf. Sporer, 1992) as well as with the number of correct details. However, these relationships were significant among re-readers only, being nearly equal in size for free reports and specific questions. It seems that recognition performance was facilitated when the descriptions that participants were allowed to re-read contained a large number of

---

**Table 6.** Mean weighted effect sizes for the re-reading effect (re-reading versus no description control group) for overall correct identification decisions, correct identifications in target-present lineups and correct rejections in target-absent lineups.

<table>
<thead>
<tr>
<th>DV</th>
<th>k</th>
<th>N</th>
<th>LOR</th>
<th>OR</th>
<th>95% CI</th>
<th>Z</th>
<th>pZ</th>
<th>Q</th>
<th>pQ</th>
<th>I²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall correct identification decisions</td>
<td>4</td>
<td>352</td>
<td>.764</td>
<td>2.15</td>
<td>1.34</td>
<td>3.44</td>
<td>3.179</td>
<td>.001</td>
<td>6.924</td>
<td>.074</td>
</tr>
<tr>
<td>Correct identifications in TP lineups</td>
<td>4</td>
<td>176</td>
<td>.668</td>
<td>1.95</td>
<td>.99</td>
<td>3.84</td>
<td>1.932</td>
<td>.053</td>
<td>3.937</td>
<td>.268</td>
</tr>
<tr>
<td>Correct rejections in TA lineups</td>
<td>4</td>
<td>176</td>
<td>1.008</td>
<td>2.74</td>
<td>1.27</td>
<td>5.92</td>
<td>2.564</td>
<td>.010</td>
<td>3.837</td>
<td>.280</td>
</tr>
</tbody>
</table>

Notes: LOR = logged odds ratio; OR = odds ratio; CI = confidence interval; LL = lower limit; UL = upper limit; Z = significance test; pZ = significance level for Z-test; Q = heterogeneity test statistic; pQ = significance level for Q-test; I² = indicator of heterogeneity.
(correct) details. In contrast, in the description-only group, none of the relationships were significant. However, relationships with specific questions in the description only group were comparable in magnitude to those in the re-reading group, although not significant due to the smaller sample sizes.

But how could identification accuracy be related to the quantity of the re-read description for the perpetrator, while it was independent of description accuracy? As a possible explanation we assume that the mere number of (correct) details was decisive, irrespective of the presence of false details or the overall description accuracy. Hence, re-reading may have been beneficial not via directly activating the memory for the true target’s physical appearance, but by activating the associative network (Anderson, 1983) in which the face of the perpetrator was embedded, and thus providing participants a sound basis for the identification task.

Although these analyses are correlational, and thus, do not allow causal conclusions, natural variations in description quantity and accuracy, which often occur in real cases (Sporer, 1996) seem to be associated with the benefits of re-reading.

Correlations between identification accuracy and number of details as well as number of correct details were moderate in size ($r_{pb} > .30$) after the longer delay but tended to be smaller after a delay of 2 days (but sample sizes are too small to test for differences in correlations). The results are in line with the assumption that a large number of initially recalled descriptors might be more beneficial for identifications after longer delays (cf. Butler & Roediger, 2007). We encourage future researchers to use even longer retention intervals to test this.

**Quality of description ratings**

As the true accuracy of a given description is not known in actual criminal cases, we additionally rated description quality given in free reports in terms of precision, specificity and informativeness. For re-readers, all of the quality ratings as well as the composite mean rating were positively associated with identification accuracy. Thus, the current results confirm our expectation that the beneficial effect of re-reading on identification performance is associated with the quality of the re-read descriptions.

The present results are consistent with findings from police records by Valentine et al. (2003), who observed positive associations between identification of the suspect (which may or may not have been the perpetrator) and description completeness, which is vaguely comparable to our "precision" variable.

**Conclusions and practical implications**

This research provides a new look at the role of person descriptions and identifications. We replicated across two ecologically valid delays that the practice of having witnesses re-read their own prior descriptions showed better identification performance in both TP and TA lineups. Re-reading one’s description is an easily applicable system variable (Wells, 1978) that does not require any additional procedures, training or resources. Our results also contradict the VOE that has not only “declined” over decades (Schooler, 2011) but may be reversed if examined in situations representative of real crimes. Consequently, the initially cited proscription made by the Police and Criminal Evidence Act (1984) not to show the witness any descriptions of the suspect prior to the identification task appears inappropriate in light of our data.

**Notes**

1. Sample sizes of Alguna et al.’s (2014) two replication studies varied due to an error in the initial experimental protocol. The replication of the traditional verbal overshadowing condition (cf. Experiment 1; Schooler & Engstler-Schooler, 1990) was conducted as a follow-up experiment and consequently only 22 of the initially participating 31 laboratories completed this experiment.
2. We are indebted to an anonymous reviewer to direct us to pointing out the testing effect as a possible explanation for our results.
3. Odds ratios ($OR > 1$) illustrate higher observed frequencies for the context reinstatement condition compared to a particular control group. Odds ratios for the results of Cutler et al.’s (1987) and Cutler et al.’s (1986) studies were converted from the reported $d$ values (for the exact formulae, see Borenstein, 2009).
4. The verbatim definitions used are available from the authors.
5. Unfortunately, 27 free report and specific questions descriptions were lost in the process of moving offices. Thus, only 101 descriptions could be used for these analyses.
6. Effects of verbalisation and re-reading on choosing rates and confidence are available from the first or second author on request. We do encourage future researchers to investigate whether the confidence-accuracy relationship is affected by re-reading, or more generally, by verbalisation of the target’s face.

**Table 7.** Mean weighted effect sizes for the re-reading effect (re-reading versus description-only group) for overall correct identification decisions, correct identifications in target-present lineups and correct rejections in target-absent lineups.

<table>
<thead>
<tr>
<th>DV</th>
<th>$k$</th>
<th>$N$</th>
<th>$LOR$</th>
<th>$OR$</th>
<th>LL</th>
<th>UL</th>
<th>$Z$</th>
<th>$p_z$</th>
<th>$Q$</th>
<th>$p_Q$</th>
<th>$I^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall correct identification decisions</td>
<td>6</td>
<td>919</td>
<td>.144</td>
<td>1.15</td>
<td>.88</td>
<td>1.52</td>
<td>1.031</td>
<td>.302</td>
<td>5.167</td>
<td>.396</td>
<td>3.225</td>
</tr>
<tr>
<td>Correct identifications in TP lineups</td>
<td>5</td>
<td>315</td>
<td>-2.62</td>
<td>.77</td>
<td>.48</td>
<td>1.24</td>
<td>1.072</td>
<td>.284</td>
<td>7.315</td>
<td>.120</td>
<td>45.320</td>
</tr>
<tr>
<td>Correct rejections in TA lineups</td>
<td>5</td>
<td>315</td>
<td>.585</td>
<td>1.79</td>
<td>1.08</td>
<td>2.98</td>
<td>2.254</td>
<td>.024</td>
<td>.432</td>
<td>.980</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Notes: $LOR =$ logged odds ratio; $OR =$ odds ratio; CI = confidence interval; LL = lower limit; UL = upper limit; $Z =$ significance test; $p_z =$ significance level for $Z$-test; $Q =$ heterogeneity test statistic; $p_Q =$ significance level for $Q$-test; $I^2 =$ indicator of heterogeneity.
Acknowledgements
The research was conducted at the University of Giessen, Germany, and at the Arizona State University, Phoenix. We would like to thank Collin G. Tredoux for commenting on earlier versions of this manuscript. Special thanks also to Reinhard Bentrup for helping us to create the ROC curves in R. The authors acknowledge that they had no financial interests or benefits that arise from the direct application of their research.

Disclosure statement
No potential conflict of interest was reported by the authors.

References
Appendix

Now we will ask you some specific questions regarding the outer appearance of the perpetrator. Please answer these questions on answer sheet B2. If you have already answered one of the questions in your previous report, please answer it again. If you don’t know the answer to one of the questions, write “don’t know” in the corresponding line.

What age do you estimate the perpetrator was?
What size do you estimate the perpetrator was?
Please describe the figure of the perpetrator in detail!
Please describe the clothes of the perpetrator in detail!
Please describe the color of the perpetrator’s hair in detail!
Please describe the perpetrator’s hairdo in detail!
Please describe the shape of the perpetrator’s face!
Did you notice any special features on the perpetrator?
If so, which?
Did the perpetrator wear a headdress? If so, which?
Did the perpetrator wear glasses? If so, what did they look like?
Did the perpetrator have a beard? If so, what did it look like?
Did the perpetrator speak in a certain dialect or did he have an accent? If so, which?