Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Experimental Social Psychology 41 (2005) 36–47
Journal of Experimental Social Psychology
www.elsevier.com/locate/jesp

Mental control and visual illusions: Toward explaining race-biased weapon misidentifications☆

B. Keith Payne,[a,*] Yujiro Shimizu,[b] and Larry L. Jacoby[b]

[a] *Department of Psychology, The Ohio State University, Columbus, OH 43210, USA*
[b] *Washington University, USA*

Received 28 July 2003; revised 2 May 2004
Available online 5 June 2004

Abstract

People are biased to misidentify harmless objects as weapons when the objects are associated with African Americans (Payne, 2001). Two studies examined the processes underlying this bias. The illusory perception hypothesis argues that stereotypes alter the subjective construal of the object. In contrast, the executive failure hypothesis argues that even when perception of the item is intact, misidentifications can result from failures to control responses. Immediately after making an error, participants were able to accurately express that they had made a mistake via confidence ratings (Experiment 1) and by correcting their judgment (Experiment 2). Subjective confidence judgments were extremely well calibrated to accuracy, and participants virtually never believed their own mistakes. Conditions likely to create errors through both illusions and control failures are discussed.

Keywords: Implicit; Prejudice; Stereotyping; Automatic; Controlled; Executive; Illusion; Weapon; Shooting

People tend to misidentify harmless objects as weapons when the objects are associated with African Americans. This bias was reported by Payne (2001) and subsequently found by two other laboratories (Correll, Park, Judd, & Wittenbrink, 2002; Greenwald, Oakes, & Hoffman, 2003). The weapon bias occurs regardless of whether the stimuli are complex scenes (Correll et al., 2002; Greenwald et al., 2003) or simple photographs of faces and objects (Payne, 2001). It occurs whether the judgment is framed as a perceptual gun/tool classification (Payne, 2001) or as a behavioral shoot/do not-shoot decision (Correll et al., 2002; Greenwald et al., 2003). Finally, the bias occurs among African American participants as well as among White Americans (Correll et al., 2002).

These findings are of theoretical and practical interest because they link recent advances in implicit prejudice research with significant socio-political events, such as the mistaken shooting deaths of Amadou Diallo and Timothy Thomas. In both of these well-publicized incidents, the victims were young Black men who were killed when police officers mistakenly responded as if they were armed. These two cases sparked peaceful protests and violent riots, civil suits and criminal trials, and new legislation aimed at reducing race biases in police practices. Because of the potential implications of research showing systematic race biases in weapon identification, it is important to develop a detailed theoretical understanding of the processes underlying this bias.

Two major classes of process explanations might account for the race bias. The first is that misidentifications result from distorted perceptions, and the second is that they result from failures to control one's responses. Both accounts are plausible, and have precedent in other literatures. The goal of the present research is to directly evaluate these two accounts.

The *illusory perception hypothesis* argues that participants misperceive the objects in what amounts to a

☆ We would like to thank Elizabeth Marsh and Alan Lambert for helpful feedback on earlier drafts of this paper. Thanks also to Jim Sherman, Josh Correll, and two anonymous reviewers for helpful suggestions on an earlier draft of the paper. Thanks finally to the Social Cognition Research Group at Ohio State University for helping refine the ideas developed here.

* Corresponding author.
E-mail address: payne.265@osu.edu (B.K. Payne).

0022-1031/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jesp.2004.05.001

perceptual illusion. This explanation assumes that perceivers use stereotypes as cues to resolve perceptual ambiguity, thus influencing their construal of the object. This hypothesis is consistent with the explanation behind scores of heavily studied perceptual illusions. The well-known Mueller–Lyer illusion illustrated in Fig. 1 serves as an example. Here, the vertical line on the left appears longer than the one on the right. But in reality the line on the right is very slightly longer, as a ruler will confirm. Most viewers of this illusion find it compelling and will assert with high confidence that the left line looks longer.

One prominent explanation for this illusion holds that the visual system compensates for the fact that objects viewed in depth change apparent size depending on how far away they are. In most situations in daily life, lines flanked by concave angles like the left arrow recede away from us in depth (like the corner of a room or the far leg of a table). In contrast, lines flanked by convex angles like the right arrow jut out toward us (like the near corner of a table). The visual system compensates for these expected differences in size that co-occur with depth cues by interpreting the left line as longer than the right line.

By analogy, the illusory perception hypothesis holds that race cues provide a context which the mind uses to "adjust" or "fill in" aspects of the scene as it transforms raw sensation into the perception of a meaningful object. The visual assumption revealed by the race bias would be that items associated with African Americans are likely to be dangerous. The illusory perception hypothesis suggests that the mind incorporates these top-down assumptions by interpreting some objects as weapons when they are paired with Black racial cues.

The illusory perception hypothesis is consistent with social psychological perspectives that emphasize subjective construal as a major mechanism responsible for creating biases. Bruner's (1957) seminal work on "perceptual readiness" argued that the cognitive accessibility of potential categories determined how a stimulus was categorized, and thus how it was perceived. According to the constructivist approach fostered by the "New Look" movement, priming with African American faces should have the effect of making the weapon category more accessible. This heightened accessibility should cause harmless items to be misperceived as weapons some portion of the time, creating the race bias in question (see Payne, Jacoby, & Lambert, in press for a related discussion).

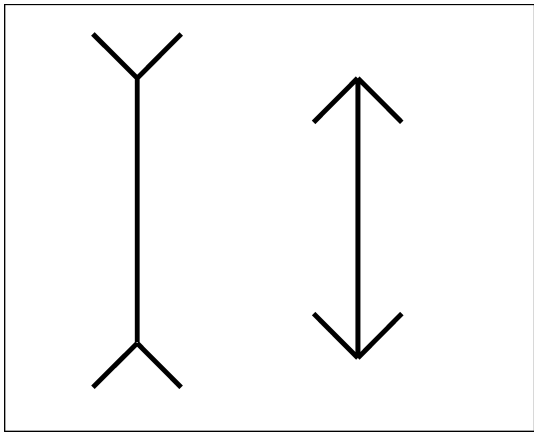

Fig. 1. The Mueller–Lyer illusion.

Research on race biases in weapon identification to date has used language that is generally consistent with this interpretation. For example, Payne (2001) referred to the bias as "misperceiving a weapon," (p. 181). Similarly, Correll et al. (2002) described the stereotype bias as an effect that "can act as a schema to influence perceptions of an ambiguously threatening target" (p. 1325). Finally, Greenwald and colleagues (2003) emphasized the "perceptual ability to discriminate a weapon from a harmless object" (p. 405). This constructivist approach with its emphasis on subjective construal seems to be a common perspective from which researchers interpret this phenomenon. One way to think about the illusory perception hypothesis is by asking about the subjective reaction of a person who has just mistakenly "fired" at an unarmed suspect. Does the person immediately regret the snap decision that he or she knows to be a mistake? Or does the person firmly believe that they saw a gun? A false perception presumably would lead to the second reaction.

In contrast to the illusory perception account, the *executive failure hypothesis* argues that errors can occur even when perceptions of the objects are intact. The problem is that people fail to execute their actions as they intend. Executive control describes an ability to plan and carry out selective behaviors in a way that follows one's goals. In many cases, this includes the need to override responses that are highly activated or well-learned but inappropriate. Because "selection" is an inherently relative concept, it is necessarily accompanied by the need to inhibit or suppress other potentially distracting information. Specifically, race stereotypes are expected to be activated by the race cues in the weapon identification task. Thus two different streams of information are available as bases for making responses: accessible stereotypes and the actual target item. Executive control performs a gating function, selectively allowing the appropriate information to control actions, while averting the influence of activated but inappropriate information. The executive failure hypothesis suggests that in a lack of coordination between eye, brain, and hand, participants' *actions* are systematically biased even though they may be aware that they have made an error.

Extreme examples of executive failure can be seen among people with brain damage to areas in the pre-

frontal cortex, a region strongly implicated in strategic planning and control of cognition and action. One well-known result of frontal damage is *perseveration*, in which patients persist in performing a well-learned activity after it is no longer appropriate to do so. Kimberg and colleagues described a striking example in which patients with frontal lobe damage learned a particular rule by which to sort cards on the Wisconsin Card Sorting Task (Kimberg, D'Esposito, & Farah, 1997). On the next block of the task, the rule was changed, so that the patients were supposed to sort cards according to a criterion different from their well-learned rule. The patients were unable to override the previously learned rule, and so continued sorting according to the out-dated rule (a typical case of perseveration). Intriguingly, the patients demonstrated knowledge of the correct rule. Nevertheless, they continued the incorrect sorting behavior even as they correctly articulated the new rule and how they should be sorting differently! The executive failure hypothesis suggests that when people display race bias in identifying weapons, they might be aware of the true identity of the items in the same way as in the above example, frontal-lobe patients were aware of the correct rule.

Of course, most individuals falling prey to the race bias do not suffer from brain damage. However, many everyday situations in which people are rushed, distracted, tired, or anxious can interfere with normal executive functioning, leading to failures of intentional control. Consistent with this reasoning, the race bias in the weapon identification task has been shown to decrease linearly with processing time (Payne, 2001; Payne, Lambert, & Jacoby, 2002). Because controlled processes require more time to execute than automatic processes (Shiffrin & Schneider, 1977), processing time is one important criterion for establishing the role of executive control. This bit of evidence suggests that executive control is important for understanding performance in weapon identifications. However, that research manipulated processing time either between participants or between different experimental blocks. Because it did not examine participants' responses to the same item over time, it does not directly speak to whether illusory perceptions or executive failures best characterize weapon misidentifications.

Previous research has at times suggested interpretations consistent with an executive failure hypothesis. For instance, Correll et al. (2002) compared the race bias to a Stroop effect, in which the automatic process of word reading interferes with the controlled process of color naming (this analogy will be fleshed out in more detail in the following section). Payne and colleagues (2001; Payne et al., 2002; Lambert et al., 2003) investigated the role of executive control by adopting Jacoby's (1991) process dissociation procedure to separate automatic and controlled influences contributing to the bias.

The most direct evidence for the engagement of executive control processes may be found in a study by Amodio et al. (2004). These researchers adapted the sequential priming procedure used by Payne (2001). Event-related potentials (*ERP*'s) were used to measure neural activity while participants performed the weapon identification task. Specifically, this study measured a particular ERP component known as *error-related negativity* (ERN). The ERN is a negative deflection in the ERP wave that is specifically sensitive to conflict detection. The ERN is found when people process information that creates conflicts which need to be resolved, as in cases of response competition. According to a recent neuropsychological model of executive control, two neural systems act together to exert control in conflict situations (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick, Nystrom, Fissel, Carter, & Cohen, 1999). One system (associated with the anterior cingulate cortex) detects conflicts and signals the need for executive control, while the other system (associated with prefrontal regions) implements the intended response.

In the Amodio et al. (2004) study, the ERN was associated most strongly with the stereotype-consistent error of misidentifying a tool as a gun in the context of a Black prime. What is critical about these findings for the present purposes is the remarkably early timing of this conflict detection mechanism. Although the ERN signal peaked at about the same time responses were made, the increase in the ERN signal began well *before* participants ever made a response. The neurocognitive model of executive control advocated by these authors suggests that the ERN reflects a conflict-monitoring process which signals the need for control processes to resolve the conflict. If so, then we might expect an association between the ERN signal and measures of response control. In fact, that is just what the data showed—the ERN signal was significantly related to process dissociation estimates of control.

Although this evidence is certainly consistent with the executive failure hypothesis, it is still unknown whether participants are aware of their errors when misidentifying weapons. In the present study we build upon previous research by examining whether participants can distinguish their correct responses from errors. The illusory perception hypothesis predicts that participants should be unable to distinguish when they are falling prey to an illusion and when they are not. In contrast, the executive failure hypothesis predicts that participants will make stereotype-congruent errors under time pressure, but that they should know when they do so. Of course, these two explanations are not mutually exclusive. Both could contribute to mistaken weapon judgments. After comparing the ability of these explanations to account for results in the weapon identification paradigm, we consider in the general discussion the conditions under which each account is most likely to apply.

Overview of the present research

In this research we used the sequential priming procedure from Payne (2001). Other studies (Correll et al., 2002; Greenwald, Oakes, & Hoffman, 2003) have found similar results using other procedures, although this does not guarantee that the results are all mediated in the same way. We chose this paradigm as a starting point for clarifying the processes underpinning the race bias, but it will be important to test whether similar processes account for the findings using other procedures as well.

In the present studies we asked participants to respond twice to each item: once under time pressure, when executive control is limited, and once at their own pace, when participants have ample opportunity to control their responses. We took care to control the amount of viewing time, so that the stimuli were removed from view before either response was made. With fast responding, both illusory perceptions and executive failures could account for any bias observed. Critically, when participants respond more slowly to the exact same item, those responses should still reflect any perceptual illusions but should not reflect executive failures. That is because slow responding is expected to allow participants a very high level of opportunity to control their responses.

To illustrate, consider the visual illusion displayed in Fig. 1. Because the illusion affects the perception of the picture, perceivers will be influenced by the illusion regardless of whether they express their response in a hurried way, or in a more careful way. Assume we ran an experiment in which participants saw 100 such figures, half of which were illusions (the lines were actually identical) and half of which were not illusions (the line that looked longer actually was longer). Participants would have difficulty distinguishing the illusory differences from the true differences. Importantly, we would not have to speed participants' responses, distract them, or otherwise interfere with their central processing abilities to observe the illusion. If we asked participants to respond twice to each figure, once under time pressure and once at their leisure, we would expect roughly similar responses both times.

Now consider an example of a kind of error driven by executive failures. In the well-known Stroop (1935) color-naming task, participants see a series of color words printed in ink colors that are sometimes inconsistent with the identity of the word. For example, participants might see the word *RED* printed in blue ink. Participants' task is to name the ink color, not to read the word. However, because word reading is a highly learned skill, it often interferes with the intended task of ink-color naming. The typical Stroop effect shows that participants are slower and less accurate when they are naming a color incompatible with the meaning of the word. Resisting the automatic impulse to read words and enacting the intention to name colors is a difficult task requiring considerable executive control. In the example described here, participants would be fairly likely to mistakenly respond with the name of the word (''red'') rather than the color of the ink (''blue'').

However, this kind of error is not a visual illusion. Instead, it reflects a failure of executive control. Immediately after making such an error, participants can correctly express the fact that they made a mistake. If we gave participants two chances to respond to each item, they would likely make a fair number of mistakes on the hurried response (Lindsay & Jacoby, 1994). However on the second, slower response, they could correct those errors with great accuracy. The ability to correct one's mistakes is a powerful indicator of when participants believe their errors to reflect reality (as in visual illusions), versus when they know them to be false (as in executive failures).[1]

The following studies allowed participants to make two responses on each trial of the weapon identification task. In the first response, participants classified the object as a gun or not, under a response deadline. Immediately following that response, they were allowed to make a second response at their own pace. Importantly, the target objects were masked and removed from view before even the first response was made. This served two important purposes. First, it prevented participants from further viewing, so that any differences observed in the slow responses were not due simply to viewing the target item longer. Second, visual masks disrupt the formation of sensory (iconic) memory, so participants could not use these sensory representations as a basis for improving their second responses. In the first experiment, participants' second response was a confidence judgment. They rated how confident they were in the response they had just made. In the second experiment,

[1] The conceptual analogy between cognitive control in weapon identification and control in the Stroop task does not imply that the processes are identical. For example, the process dissociation model used here assumes that automatic bias influences responses only when executive control fails. In contrast, Lindsay and Jacoby (1994) developed a process model of the Stroop task that assumes that controlled color-naming drives responses only to the extent that automatic word-reading processes are inhibited. The main difference between these two models is which process is contingent on the other. Nonetheless, controlled processes in both models refer to responding in accord with intentions rather than responding on the basis of automatic processes. The authors have used multinomial modeling procedures to test which kind of model best fits the weapon identification task. Across several data sets, the former model fit somewhat better than the latter Stroop-like model, but both models sometimes showed acceptable fits to the data. Our point is not that executive control capacities are engaged identically in the two tasks, but that the Stroop task provides a good example of how executive control is needed to avoid the influence of automatic processing.

the second response was to classify the item as a weapon or a tool, a judgment identical to the first judgment.

If participants' errors are produced by a visual illusion, then their reports on the slower, second response should coincide with their initial, hurried reports. They should report confidence levels for erroneous responses comparable to those for correct responses (Experiment 1). And when they make the identification judgment twice (Experiment 2), the second judgment should be the same as the first. However, if participants' errors are produced by executive failures, then they should make a stereotype-consistent pattern of errors on the initial fast response, but not the second, slower response. Their confidence judgments should discriminate between correct and incorrect responses; and their second identification judgment should be highly accurate even when their initial judgment was incorrect.

Experiment 1

Method

Participants

Thirty-three undergraduates participated for course credit. All participants classified themselves as either White or Asian Americans. Participants were run at computer terminals in individual rooms.

Design

The stimuli were drawn from the materials used by Payne (2001). The prime stimuli included 5.3 cm × 4 cm images of four Black and four White faces. Photographs of four guns and four tools were used as targets. The design was a 2 (Prime race: Black, White) × 2 (Target: Gun, Tool) factorial design. During each trial, visual masks were shown before each prime and after each target. These visual masks consisted of a rectangle of randomly placed black and white dots.

Procedure

Participants were instructed that they would be performing two tasks on each trial. They were told that their first task was to identify a briefly presented object as a gun or tool by pressing one of two keys. Their second task was to give their subjective evaluation of how confident they were in that judgment. The experimenter explained that prior to the presentation of the object, they would see a face. They were not told to do anything with these images, but were told that they would serve as a warning signal that the critical object was about to appear.

For each trial, the sequence of events was as follows: first, a mask appeared for 500 ms, followed by a prime for 200 ms, then a target for 100 ms, and finally a mask again for 500 ms or until the participants responded, whichever came first. Participants were encouraged to identify each item quickly, and informed that they would have less than a second to make a response. If they did not respond to the object within 500 ms, a red "X" appeared to provide negative feedback.

Immediately after participants made a gun or tool response by pushing one of two response keys (with two fingers of the left hand), they were asked for their subjective confidence in their judgment on a six-point scale. The scale was anchored by values ranging from 1 (not at all certain the response was correct) to 6 (complete certainty the response was correct). It was explained to participants that because there were only two response options, they had at least a 50% chance of responding correctly even if they did not know what the item was. There was no time limit for this rating, and participants were encouraged to take as much time as they wished for the confidence rating. Responses were entered by pressing one of six keys with participants' right hands.

Participants were instructed to place two fingers from one hand over the "gun" and "tool" response keys for the first response, and to place the other hand over the row of response keys labeled for the confidence scale for the second response. After the rating was made, a "ready" prompt appeared for 1 s, after which the next trial began. Each participant completed one block of 64 practice trials, followed by 3 critical blocks of 64 trials (192 total) during the main test phase. Following this procedure, participants were fully debriefed and dismissed.

Results

The main question of interest was whether participants' confidence judgments successfully differentiated between their errors and their correct responses. If participants experienced an illusion similar to the Mueller–Lyer illusion, then their confidence judgments on stereotype-consistent error responses should be similar to their confidence judgments on correct responses. However, if participants experienced failures of executive control, then their confidence judgments on error responses should be low, whereas their confidence judgments for correct responses should be high. To address this issue, results will first be reported for the weapon identifications made under the response deadline. Following that analysis, confidence judgments will be examined.

Misidentifications

The proportion of errors in each prime and target object condition were tabulated and analyzed using a within-participants analysis of variance (ANOVA). Because the task required a two-alternative choice, an error on a tool meant calling it a gun, whereas an error on a gun meant calling it a tool. Results closely replicated

those of Payne (2001). First, there was a main effect of target object, $F(1, 32) = 4.66$, $p < .05$, indicating that participants made fewer errors for actual guns ($M = .07$) than for tools ($M = .10$). More importantly, the two-way Prime race × Target interaction was significant, indicating that participants were more likely to misidentify a tool as a gun than vice versa after a Black prime, not after a White prime, $F(1, 32) = 4.64$, $p < .05$.

Simple effects tests confirmed that the difference between errors for tools versus guns was significant after a Black prime, (M's = .11 vs. .06) $F(1, 32) = 9.51$, $p < .01$, but not after a White prime, (M's = .09 vs. .08) $F(1, 32) = 0.38$, *ns*. These results converge with the several studies that have found stereotype-consistent errors in weapon identification. Participants were prone to responding as if a harmless item was a weapon, but only in the context of a Black racial cue.

Process dissociation analysis

Next we report results using the process dissociation procedure (Jacoby, 1991; Jacoby, Toth, & Yonelinas, 1993). Process dissociation is an analytic technique that allows automatic and controlled components of task performance to be separated. In our previous work, we found that race priming, the salience of race, and racial attitudes were associated with the automatic component (Payne, 2001; Payne et al., 2002), whereas time pressure and anxiety had disruptive effects on the controlled component (Lambert et al., 2003; Payne, 2001). The controlled component reflects the ability to discriminate between guns and tools without bias from stereotypes. Our theoretical framework argues that in order to do so, people must maintain their goal to identify weapons while ignoring the interfering effect of race stereotypes. As such, the ability to discriminate between guns and tools without bias reflects the operation of executive control processes.

These assumptions, while important to our broader theoretical framework, would be circular if applied to the present question of whether weapon misidentifications reflect illusions versus executive failures. The process dissociation estimates can be conceptualized in a more theory-neutral way by referring to the "controlled" component simply as *discriminability*. As such, discriminability reflects the ability to behaviorally distinguish between weapons and non-weapons, independent of any response biases. At this level, the estimate is similar to other parameters that have been used to estimate discriminability such as sensitivity (d') in signal detection theory (Correll et al., 2002; Greenwald et al., 2003). However, it is important to note that discriminability does not necessarily mean perception. Poor discriminability can result either from perceptual failures (as in the illusory perception hypothesis) or from failures to control actions (as in the executive failure hypothesis). By treating the discriminability estimate in a theory-neutral way, the results of the present experiments can shed light on whether the "executive control" assumptions are justified.

The discriminability estimate normally ranges from 0 to 1, where 0 means no ability to distinguish weapons from non-weapons, and 1 indicates perfect accuracy. Accessibility bias refers to the influence of the race primes in making stereotype-consistent information accessible. Accessibility drives responses when discriminability fails, creating a bias to respond in stereotype-consistent ways. When the target item is counter to the stereotype (e.g., a harmless item paired with a Black person) this bias leads to stereotypical errors (false alarms). When the target happens to be consistent with the stereotype (e.g., a gun paired with a Black person) this bias leads to correct responses (hits).

Here we briefly describe how estimates of discriminability and bias are computed in this paradigm. For a more detailed treatment, see Payne et al. (in press). The pairing of White and Black primes with gun and tool target items creates experimental conditions in which the pairings are stereotypically *congruent* (Black–gun, White–tool) and *incongruent* (Black–tool, White–gun). In the congruent condition, responding based either on discriminating the target item (D) or based on the activated stereotype (A) in the absence of discriminating (1 − D) would lead to the same correct response. Thus, the probability of a correct response in the congruent condition is D + A (1 − D). In the incongruent condition, however, only responding based on discriminating the target (D) would lead to correct responses. Responding based on a stereotype-consistent accessibility bias (A) when unable to discriminate (1 − D) would lead to a stereotype-consistent error. Thus, the probability of a stereotype-congruent error in the incongruent condition is A (1 − D). Because the actual probabilities of correct and incorrect responses in each condition are known from the experiment, we can solve algebraically for estimates of D and A.

This procedure was used to compute estimates of each process in each prime condition in the present experiment. Replicating our previous work, results showed that race had no effect on discriminability (D), $F(1, 32) < 1$, *ns*. Discriminability was high in general, and it was just as high for Black primes ($M = .83$) as for White primes ($M = .83$). Estimates of accessibility bias, in contrast, showed a significant effect of Prime race, $F(1, 32) = 6.04$, $p < .05$. Accessibility bias estimates were scored so that higher numbers represent a greater propensity to respond "gun." Accessibility bias estimates were higher for Black primes ($M = .63$) than for White primes ($M = .48$). Replicating our previous findings, the effect of race on misidentifications was shown to be due to a stereotypic accessibility bias. However, accessibility bias drives responses only to the extent that discrimination fails. The high degree of discriminability

in the present experiment explains why the race bias observed in performance was relatively small on average.

The critical comparisons for the present purposes are between responses at varying levels of confidence. If errors are the result of illusions, then participants might be expected to make incorrect responses with high confidence, comparable to correct responses. However, if the errors reflect executive failures, then confidence should be well calibrated to accuracy. In the case of stereotype-consistent errors, if executive failure is involved, participants should respond incorrectly and immediately be able to express that they have done so via a low confidence judgment.

Confidence judgments

To examine the relationship between accuracy and confidence, average levels of confidence were computed as a function of whether the preceding identification response had been correct or incorrect, and whether it was a stereotype-congruent trial or a stereotype-incongruent trial. Mean confidence was then analyzed using a 2 (accuracy) $\times$ 2 (stereotype-congruency) repeated measures ANOVA. The prime $\times$ target interaction was collapsed into stereotype-congruency to avoid dropping several participants who had no errors in one of the four prime $\times$ target cells. By conducting the analysis in this way, we retained most participants, but two were still excluded because they had no errors in either the stereotype-congruent or the stereotype-incongruent condition.

Results revealed only a main effect of accuracy, $F(1, 30) = 221.18$, $p < .001$. As shown in Fig. 2, participants were highly confident after making a correct response, and not at all confident after making an error. Because the scale options ranged only from 1 to 6, the size of this relationship ($\eta^2 = .88$) suggests that participants had an impressive ability to monitor and express their own mistakes. This impressive calibration did not differ as a function of whether errors were stereotype-consistent or stereotype-inconsistent ($F < 1$ for the Stereotype-congruency $\times$ Accuracy interaction). This fact is important in comparing errors made under time pressure to the confidence judgments made afterward. First, recall from the initial analysis that participants did make more stereotype-consistent errors than stereotype-inconsistent errors. However, their confidence judgments were not "fooled" on these stereotype-consistent misidentifications. Even when making a stereotype-consistent misidentification (e.g., mistaking a wrench for a gun when associated with a Black person), participants were able to express extremely low confidence, suggesting that they were aware of their mistake.

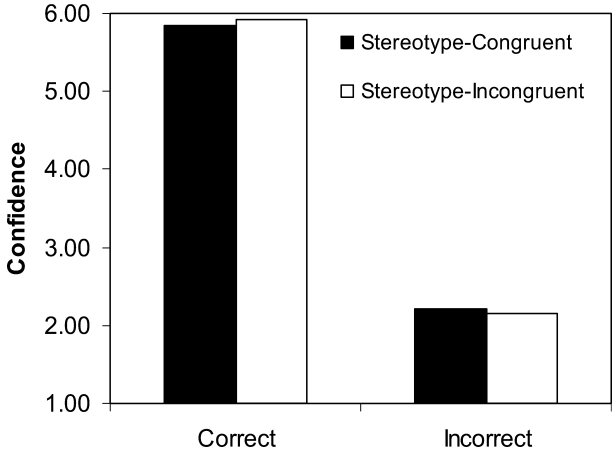

Fig. 2. Confidence ratings as a function of whether the response was correct or incorrect, and whether the trial was stereotype-congruent (e.g., Black–gun, White–tool) or stereotype-incongruent (e.g., Black–tool, White–gun). Scores have a possible range from 1 to 6.

The preceding analysis examined confidence as a function of accuracy. This answered the question of whether a correct response was likely to be given a high confidence rating, and whether an incorrect response was likely to be given a low confidence rating. An alternative way of examining the relationship between confidence and accuracy is to examine accuracy as a function of confidence. That is, given that a participant judged a particular response as highly confident, what is the likelihood that the response was, in fact, correct? To answer this question, we calculated for each participant the average confidence assigned to correct and incorrect responses. Next we used logistic regression to predict accuracy as a dichotomous outcome from the continuous confidence rating. This analysis again showed a strong relationship between confidence and accuracy, $B = 3.27$, $SEB = 1.17$, $p < .005$. The model was able to correctly classify 97% of responses as accurate or inaccurate based on the confidence rating. Finally, the Phi coefficient was .97.

To more directly investigate the relationship between confidence judgments and discriminability, the process dissociation discriminability estimate was computed for each level of confidence. Fig. 3 displays the average D estimate as a function of confidence. This analysis showed a significant relationship between confidence and the D estimate, $F(5, 160) = 112.35$, $p < .001$. The relationship was very strong, $\eta^2 = .78$. Linear trends analysis showed that most of the variance was explained by the linear trend, $F(1, 32) = 496.86$, $p < .001$, $\eta^2 = .94$. However, the cubic ($\eta^2 = .61$) and fifth-order ($\eta^2 = .22$) trends also were significant (both p's $< .005$). Note that the normal range of the D estimate is between 0 and 1, with zero reflecting chance performance. However, the lowest confidence ratings were associated with D estimates well *below* zero. This reflects the fact that, of those trials assigned a confidence level of 1, participants were wrong on the vast majority of responses. But perhaps a more intuitive way to phrase this is that when participants were wrong, they very often assigned a very low confidence rating.

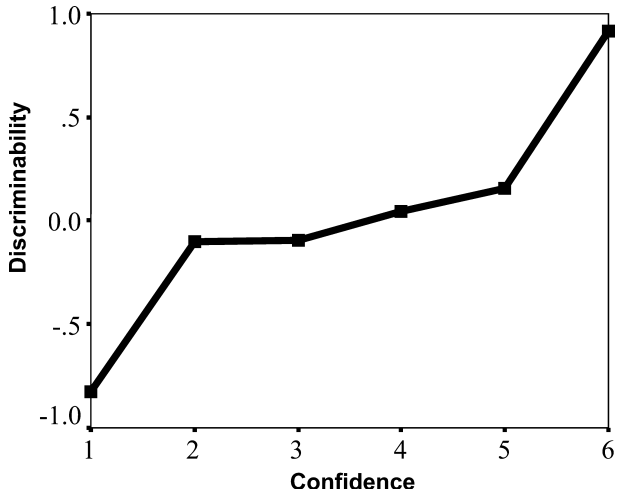

Fig. 3. Discriminability (D) estimates as a function of confidence ratings in Experiment 1.

Because the process dissociation estimate of automatic processing requires the assumption that both floor and ceiling effects are avoided, we could not calculate an unbiased estimate for the automatic component as a function of confidence. That is, at high levels of confidence most participants made no errors; at low levels of confidence, most participants made no correct responses. Nevertheless, far from being a weakness, this extremely strong relationship between confidence and the D estimate is telling. It suggests that participants had a nearly perfect ability to monitor their mistakes.

Discussion

Two aspects of these data are striking. First, participants' confidence judgments showed remarkable fidelity to their actual accuracy. When participants correctly classified items, they appropriately expressed very high confidence. On the other hand, when they misidentified an item, they appropriately expressed very low confidence. Second, this monitoring effectiveness was equally good during stereotype-consistent errors and stereotype-inconsistent errors. These data are inconsistent with the illusory perception hypothesis, which implies that when participants made stereotype-consistent errors, they believed they were accurate.

However, some questions remain about how participants were using the confidence scale. Initially we expected the lowest possible confidence level to reflect complete uncertainty (i.e., chance performance). This is why we anchored the scale with "complete uncertainty" and "complete certainty." But clearly, when participants expressed very low confidence they were not just expressing uncertainty. Instead, they were expressing certainty that they had made an error. Although participants were obviously able to use the scale to successfully express their confidence levels, participants might not all have used the scale in a uniform way.

A second limitation of these data is that the distribution of confidence responses was skewed and bimodal. Ninety-two percent of all confidence responses fell at the highest two scale levels (5 and 6). The second highest proportion of responses, 6%, fell at the lowest scale value (1), with the intermediate values ranging from .4 to 1%. The skew is to be expected, given that participants' actual accuracy was very high (mean accuracy = .91) and participants were very good at monitoring that accuracy. Although skewed distributions can have the effect of reducing statistical power, the strong relationships observed here suggest that statistical power did not pose a problem for interpreting these results. More problematic is that the somewhat bimodal distribution suggests a continuous scale might not be the best way to represent participants' experiences. Rather than a smooth range of confidence, participants might have experienced more of a dichotomy between certainty of correct responses and certainty of incorrect responses. To address these issues, and to more directly test the link between perceptions and actions in this paradigm, we conducted a second study.

Experiment 2

The design of this experiment was similar to the design of Experiment 1. The main change was that instead of expressing confidence after classifying the object, participants made a second classification response. That is, in their first response (under response deadline) participants responded by pressing the "gun" or "tool" key to identify the target object. In their second response for each trial (not under deadline) participants made the same judgment again. Participants were told that if they had made a mistake on the initial response, they could correct the error with the second response. Again, the target items were removed from view and masked after only 200 ms of viewing, before even the first response could be made.

This design offers an even more direct way to compare the illusory perception hypothesis and the executive failure hypothesis. On the one hand, if participants believe their errors to be an accurate reflection of what they have seen, then they should make similar patterns of errors on their initial rushed responses and on their second, slower responses. If, on the other hand, participants know when their actions have failed them, then they should make errors on their initial rushed responses, but not on the second responses. Specifically, we predicted that participants would make a stereotype-consistent pattern of errors on their initial responses (regardless of which hypothesis is correct). For the second set of responses, the predictions of the two hy-

potheses diverge. The illusory perception hypothesis predicts a similar *rate* of errors, and also a stereotype-consistent *pattern* of errors for the secondary responses as well. In contrast, the executive failure hypothesis predicts that participants will be highly accurate, with no influence of the race primes in the second response set.

Method

The design of this experiment was identical to the design of the first experiment, with the following exception. Instead of making a confidence judgment on the second response, participants made a second gun/tool decision on the previously viewed item. As in the first experiment, the first response was required to be made within 500 ms. If participants responded after that time interval a large red X appeared. As in the first experiment, the target stimuli were covered by a visual mask after being presented for 100 ms, which removed the stimuli from view before the first response could be made. The second response was not under time pressure. Participants were told that if they believed they had made a mistake on the first response, they could correct the error by changing their answer on the second response. Immediately following the first response, a probe appeared on the screen for the second response. The probe read "Actually gun or tool?" Participants made their responses by pressing the same keys as for the initial response (the Q key for guns and the P key for tools). The design of this experiment was a 2 (Prime race: Black or White) × 2 (Target item: Gun or Tool) × 2 (Response time: First or Second) factorial, with all factors within participants.

Thirty-three new participants recruited from the same pool as Experiment 1 participated in return for course credit. None of the participants was African American.

Results

Weapon misidentifications

Both the illusory perception hypothesis and the executive failure hypothesis predict a stereotype-consistent pattern of errors on the first response. On the second response, the illusory perception hypothesis predicts a similar pattern of stereotype-consistent errors, whereas the executive failure hypothesis predicts very few errors of any kind, and no impact of stereotypes. To test these two alternatives, we conducted a 2 (Prime race: Black, White) × 2 (Target: Gun, Tool) × 2 (Response time: First, Second) ANOVA. Fig. 4 shows the proportion of errors as a function of Prime race, Target, and Response.

The first finding of interest is that participants were near perfect in accuracy on the second response, but not the first response. This difference was reflected by the main effect of response time, $F(1, 32) = 39.04$, $p < .001$. The second important finding was that participants displayed a stereotype-consistent pattern of misidentifications on the first response, as predicted by both hypotheses. Tools were more likely to be misidentified as guns after a Black prime than a White prime, whereas guns were more likely to be misidentified as tools after a White prime than a Black prime. However, no such stereotype-consistent pattern emerged among the second responses. The differential impact of stereotypes between first and second responses was reflected by a significant three-way Prime race × Target × Response time interaction, $F(1, 32) = 8.97$, $p < .005$. Follow-up analyses showed that the two-way Prime race × Target interaction was significant for the first response, $F(1, 32) = 8.58$, $p < .006$, but not for the second response, $F(1, 32) < 1$, $p = .47$.

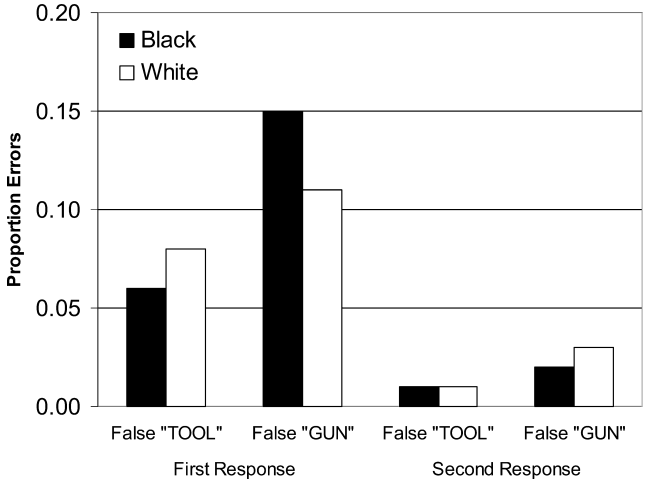

Fig. 4. Proportion of errors by type of errors (false "gun" versus false "tool" response), prime race (White versus Black) and time of response (first response versus second response) in Experiment 2.

Other significant effects included a main effect of Target item, $F(1, 32) = 7.12$, $p < .01$, indicating that participants were more accurate in identifying guns than tools. Finally, a two-way Target × Response time interaction, $F(1, 32) = 7.52$, $p < .01$ indicated that the advantage for identifying guns over tools held only in the hurried first response, $F(1, 32) = 8.07$, $p < .008$, but not in the second response, $F(1, 32) < 1$, $p = .40$.

Process dissociation estimates

The large reduction in errors from the first response to the second response suggests that participants were very good at discriminating between weapons and nonweapons, given ample time. To test the effects of response time on discriminability, D estimates were analyzed using a 2 (Response time) × 2 (Prime race) ANOVA. Results indicated that D was significantly higher on the second response than on the first response (M's = .97 vs. .77), $F(1, 32) = 39.04$, $p < .001$. Again,

Prime race had no effects on discriminability, $F(1, 32) < 1$, and did not interact with response time, $F(1, 32) = 1.21$, *ns*.

Estimates of accessibility bias could be generated for the fast response condition, but not for the slow response condition. This is because most participants had no errors at all in the slow response condition. Analyses within the fast response condition replicated the findings reported in the first experiment. Accessibility bias was higher after a Black prime than after a White prime (M's = .66 vs. .52), $F(1, 32) = 9.07$, $p < .005$. This result confirms the notion that race stereotypes were activated by the prime and thus participants had a source of bias that needed to be overcome by the later response.

Discussion

The stereotype-consistent pattern of misidentifications that has been found in several studies was replicated in the fast-response condition. But when participants responded more slowly to the same item a moment later, all traces of the stereotype bias were gone. Participants' "second thought" responses showed very high accuracy. If participants had incorrectly perceived the target items because of the race primes, we would have expected to see their errors repeated on the second set of responses. These data suggest that participants were very good at knowing when they had made a mistake. These results support the executive failure hypothesis, and provide no support for the illusory perception hypothesis, at least as it has been articulated thus far. There may, however, be other versions of the illusory perception hypothesis that are consistent with these findings. We turn to consideration of these matters in the General discussion.

General discussion

We might refine the illusory perception hypothesis by breaking it into two different hypotheses, each operating at different time scales. The version we have been focused on might be called the *enduring illusion hypothesis*, because it takes as its time scale a span of seconds. It asks whether a person, having made a false alarm, can accurately tell you that he or she has done so a few seconds later.

The second version might be called the *fleeting illusion hypothesis*, because its time scale is a matter of milliseconds. This hypothesis argues that participants experienced a brief illusion for a few hundred milliseconds after the stimulus had been presented. Participants' fast-deadline responses reflected this illusion. After about 500 ms however, the illusion faded, and participants were able to discriminate between items nearly perfectly by their second response. Although the present studies used visual masks to disrupt further viewing of the target items, the masks do not guarantee that no further visual processing took place on the representation that was initially formed.

The fleeting illusion hypothesis seems difficult to overturn decisively using behavioral responses, reminiscent of arguments that have been made against the existence of unconscious (subliminal) perception or priming. Subliminal perception is typically demonstrated by showing that participants who have been primed show behavioral influences of the prime without being able to report awareness of it. For example, participants primed with *lion* might be faster to identify the word *tiger*, while later being unable to report the nature of the prime word. One criticism of unconscious perception is that participants might have been momentarily aware of *lion*, but that the representation was obliterated when the next stimulus was presented. Therefore, "unconscious perception" might rather be a case of conscious perception with a rapid problem of memory.

Our experimental results present the inverse of the unconscious perception problem. The critic of unconscious perception argues that participants in such experiments have an accurate, conscious perception that is quickly erased or distorted. Conversely, the fleeting illusion hypothesis argues that participants misidentifying weapons have a distorted perception that is quickly restored. In both cases, a skeptical reader might find the delayed report untrustworthy in describing what the person perceived milliseconds before. If participants' subjective reports are not taken as authoritative about what they have been aware of, it becomes difficult to be certain that no fleeting illusion could have occurred.[2]

Although the fleeting illusion hypothesis cannot be ruled out, the present results are clearly inconsistent with the enduring illusion account. Unlike the experience of an illusion, such as the Mueller–Lyer illusion in Fig. 1, participants did not assert with confidence that harmless objects looked like guns. And when given a "second chance" to express their judgment, they were highly accurate and not biased by race cues. These results, therefore, help to constrain the range of mechanisms that are likely to account for misperceptions of weapons under time pressure.

[2] The philosopher Daniel Dennett (1991) has gone so far as to argue that at such millisecond time scales there is no way, even in principle, to distinguish when people become aware of something or cease to be aware of something. The idea is that because the brain processes information in modules distributed in different locations across the cortex, often at different speeds, the information available to one part of the brain is often different than that available to other parts of the brain. Because there is no one "place" that defines what a person is aware of, there is no fact of the matter at any given point in time. We mention this theory to highlight the intricacies of determining conscious awareness at very short time scales.

Our data converge nicely with the results of another recent study that suggest executive control is important in weapon misidentifications. Govorun and Payne (2004) found that process dissociation estimates of control, but not estimates of automatic bias, correlated with individual differences in working memory capacity and performance on the Stroop color-naming task. The conceptual parallel between cognitive control in weapon identification and the Stroop task was discussed previously. The correlation with working memory is noteworthy because working memory has commonly been used as a marker of central executive processes (Baddeley, 1986; Kane, Bleckley, Conway, & Engle, 2001).

The weapon identification task, the Stroop task, and the working memory task all share the theoretically interesting need to select and maintain goal-appropriate information while suppressing interference from potentially distracting information. Together with the ERP results described above from Amodio et al. (2004), these patterns of results suggest a general capacity for executive control may underpin these very different kinds of tasks. Neither version of the illusory perception hypothesis can explain why individuals with poor attentional control or poor working memory might suffer greater illusions. The executive failure hypothesis, on the other hand, makes the specific prediction that measures of executive control should predict errors on the weapon identification task.

Scope and boundary conditions

The preceding arguments should not be taken to suggest that illusory perceptions do not occur, or that they can never explain race biases in weapon identification. Instead, the present studies aim to explain the race biases previously observed in the sequential priming version of the weapon identification task. Almost certainly, errors in actual police confrontations are more complex that this laboratory model. Nonetheless, the model can be useful in understanding the basic processes contributing to such errors. When are mistaken shootings likely to result from executive failures, and when might they result from false perceptions? We speculate that the answer depends on two factors: opportunity to control one's responses and visual ambiguity. The first has been discussed at length; factors such as time pressure or distraction that interfere with control processes should make executive failures more likely. The second factor, visual ambiguity, may be equally important.

The target items we used in the present study were not very ambiguous. Given enough time to respond, participants could identify them virtually perfectly. However, poor lighting conditions and partially obscured objects might in some cases present the observer with a truly ambiguous object. It is well-known that schemas such as stereotypes are most likely to bias construal of an object or situation when it is ambiguous, and hence can plausibly be interpreted in multiple ways (Bruner, 1957).

We suggest that the relative contribution of executive failures and visual illusions can be understood by considering a 2×2 matrix in which processing constraints (e.g., time pressure) and visual ambiguity are orthogonally crossed. The high time pressure/high ambiguity cell is likely to produce errors from both executive failures and false perceptions. Unfortunately, this is probably the cell most descriptive of many actual police confrontations. The high time pressure/low ambiguity cell is likely to generate errors due to executive failure, but not from false perception. This cell is most representative of the fast-response condition in the present experiments. The low time pressure/low visual ambiguity cell is likely to produce extremely high accuracy, because neither "risk factor" for errors is present. This cell is most representative of the slow response condition in the present experiments. Finally, the low time pressure/high ambiguity cell is likely to generate errors due exclusively to biased construal. To our knowledge this hypothesis has not been tested, although it is a direction we are currently pursuing.

A related question is whether the procedures used by Correll and colleagues or Greenwald and colleagues may be more sensitive than the sequential priming procedure to illusions versus executive failures. Those procedures have used more complex stimuli in which the target item is seen in the hand of the Black or White suspect. Research showing biased error rates in those studies has required or encouraged fast responding, suggesting that executive failures may be important to those findings as well. However, future research might determine whether those conditions are also more conducive to biased perceptions.

Conclusion

We found evidence that people know when they misidentify weapons due to fast responding in the sequential priming procedure. Although their responses were biased by race, their perceptions were not necessarily distorted. Stereotypes may affect participants' actions when executive functions fail, independent of subjective construal. This view highlights the potential role of executive control processes in understanding when automatic reactions translate into behavioral biases.

References

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the

detection of unintentional race bias. *Psychological Science, 15*, 88–93.

Baddeley, A. D. (1986). *Working memory*. London: Oxford University Press.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624–652.

Botvinick, M. M., Nystrom, L. E., Fissel, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402*, 179–181.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64*, 123–152.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using race to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314–1329.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little Brown & Company.

Govorun, O., & Payne, B. K. (2004). Ego depletion and prejudice: Separating automatic and controlled components. Under review.

Greenwald, A. G., Oakes, M. A., & Hoffman, H. G. (2003). Targets of discrimination: Effects of race on responses to weapon holders. *Journal of Experimental Social Psychology, 39*, 399–405.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541.

Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General, 122*, 139–154.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*, 169–183.

Kimberg, D. Y., D'Esposito, M., & Farah, M. J. (1997). Cognitive functions in the prefrontal cortex—Working memory and executive control. *Current Directions in Psychological Science, 6*, 185–192.

Lambert, A. J., Payne, B. K., Jacoby, L. L., Shaffer, L. M., Chasteen, A. L., & Khan, S. K. (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology, 84*, 277–295.

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 219–234.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192.

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology, 38*, 384–396.

Payne, B. K., Jacoby, L. L., & Lambert, A. J. (in press). Attitudes as accessibility bias: Dissociating automatic and controlled processes. In J. S. Uleman, J. A. Bargh, & R. Hassin (Eds.). *The New Unconscious*. Oxford.

Shiffrin, R., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.