Yamaguchi, M., & Beattie, G. (in press). The role of explicit categorization in the Implicit Association Test. *Journal of Experimental Psychology: General*.

The Role of Explicit Categorization in the Implicit Association Test

Motonori Yamaguchi[12] and Geoffrey Beattie[2]

[1]University of Essex, Colchester, UK

[2]Edge Hill University, Ormskirk, UK

Author Notes

Motonori Yamaguchi, Department of Psychology, University of Essex, Colchester, the United Kingdom; Geoffrey Beattie, Department of Psychology, Edge Hill University, Ormskirk, the United Kingdom. The present study was supported in part by the Research Investment Fund from Edge Hill University. We thank David Lilley for his help in preparing the experimental materials and collecting the data. We also thank Jan De Houwer, Christoph Klauer, and an anonymous reviewer for constructive comments on an earlier version of the manuscript. The original experimental data and the stimuli used in the present experiments are available from the Open Science Framework project page (https://osf.io/gr84p/).

Correspondence concerning this article should be addressed to Motonori Yamaguchi, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK. E-mail: motonori.yamaguchi@essex.ac.uk.

Abstract

The present study investigated how task-irrelevant attributes of a stimulus affected responses in a multi-attribute version of the Implicit Association Test (IAT). In Experiment 1, participants categorized images of Black and White male and female individuals on the basis of either race or gender. Both the race and gender of the individuals affected task performance regardless of which attribute was currently relevant to performing the task, yielding the IAT effects on both attributes. However, the influences of a task-irrelevant attribute depended on whether the task-relevant attribute was categorized compatibly or incompatibly with the underlying implicit biases. These results suggest that individuals are still categorized implicitly based on task-irrelevant social attributes and that the explicit categorization required in the standard IAT has a considerable impact on implicit social biases. Experiment 2 considered a third, non-social attribute (the color of the picture frame) and reproduced task-irrelevant IAT effects and their dependence on explicit categorization. However, Experiments 3 and 4 suggested that the task-irrelevant IAT effects based on social attributes are determined by whether the task-relevant attribute is a social or non-social attribute. The results raise fundamental questions about the basic assumptions underpinning the interpretations of the results from the IAT.

Keywords: Implicit association test; implicit attitude; bias modification; automatic processes.

It is widely accepted in psychology that attitudes shape perceptions and actions in many different social contexts (Allport, 1935; Beattie, 2010; Sarnoff, 1960; Thorndike, 1920; Thurstone, 1928). In everyday social life, people have a tendency to evaluate others based on personal and social attributes that are not relevant to the matters under consideration. For example, attributes of a person, like their race, gender, or socioeconomic status, can influence other people's judgments of them on a number of dimensions, including their perceived suitability for an advertised post (Beattie, Cohen, & McGuire, 2013), their perceived guilt or the seriousness of their crime in a courtroom setting (Downs & Lyons, 1991; Porter, ten Brinke, & Gustaw, 1991), or even the perceived hostility of their facial expressions or actions (Devine, 1989). However, the social attitudes that give rise to these judgments are not always explicit to those who are actually making the judgments, and the influence of these social attitudes can go largely unnoticed (Greenwald & Banaji, 1995). These have been termed 'implicit' attitudes and have been the subject of considerable research in many different areas of psychology over years (Beattie, 2013). A range of techniques have been developed to measure such implicit attitudes (Bar-Anan & Nosek, 2014; De Houwer & Eelen, 1998; De Houwer, 2003; De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Fazio & Olson, 2003; Koole, Dijksterhuis, & van Knippenberg, 2001; Nosek & Banaji, 2001; Nosek, Hawkins, & Frazier, 2011; Olson & Fazio, 2003), of which the *Implicit Association Test* (IAT; Greenwald, McGhee & Schwartz, 1998) has been the most popular and most influential method in the last two decades.

Although the IAT has been a very popular method to test 'implicit' biases toward specific attributes of individuals, the task is designed to force participants to categorize individuals based on the very attributes that are tested. The purpose of the present study was to examine whether the IAT effect can occur based on attributes that are completely irrelevant to the task in hand. To

this end, we used versions of the IAT that involved one or more task-irrelevant attributes, which we call a *multi-attributes IAT*, or *m-IAT*. In what follows, we first describe the design of the IAT and introduce the relevant terms that we adopted in the present study. Next, we point out how the IAT forces explicit categorization of individuals against the attribute in question and raises the fundamental issue of the implicitness of biases that it is meant to measures. We then consider why explicit categorization may have an impact on the expression of implicit biases in the m-IAT. Four experiments addressed the questions of whether social biases are still observed when the attribute (e.g. race or gender) is irrelevant to the task and of how one's intention to categorize an attribute might affect the influences of other biases. Finally, the theoretical implications, and the societal significance, of the present set of findings are discussed.

Implicit Association Test

The IAT assesses underlying associations between attributes of stimuli (e.g., the names of typical White and Black individuals) and evaluative qualities (e.g., 'pleasant' vs. 'unpleasant') by testing how quickly people can sort the stimuli into specific categories that are associated with an evaluative quality or a stereotype. It measures response time (RT) of categorization (and sometimes error rate) and typically shows a bias with certain types of categorization rules (e.g., 'White names' to 'good') being quicker than others (e.g., 'Black names' to 'good'). Many authors have concluded that this demonstrates an implicit bias with potentially highly significant societal consequences (e.g., Beattie, 2013). Although the IAT has been a popular tool in many domains of implicit bias and attitudes, several issues have been raised that question its validity for measuring possible social biases (e.g., Arkes & Tetlock, 2004; Blanton, Jaccard, Strauts, Mitchell, & Tetlock, 2015; Blanton, Jaccard, Gonzales, & Christie, 2006; Fiedler, Messner, & Bluemke, 2006; Gawronski & Bodenhausen, 2011; Nosek & Sriram, 2007).

In the standard procedure, the IAT consists of two types of categorization tasks, which are often called ‘attribute’ and ‘target’ tasks in the IAT literature but we call the *inducer task* and the *diagnostic task* (the terminology adopted from previous studies of compatibility effects; e.g., Notebaert, Gevers, Verguts, & Fias, 2006) to express their functional roles in the procedure more accurately (see also De Houwer, 2003, for a similar distinction between two mixed tasks). The *inducer* task induces evaluative qualities (e.g. ‘good’ vs. ‘bad’) in non-evaluative responses (left and right key presses), and the *diagnostic* task is used to identify biases against specific categories (e.g., Black and White individuals when paired with the evaluative concepts). The performance of categorization in the diagnostic task is the primary basis of the measure of implicit attitudes. In the original design (Greenwald et al., 1998), the inducer task required categorizing a set of words into ‘pleasant’ or ‘unpleasant’ based on their meanings by pressing one of two response keys. The diagnostic task required categorizing another set of words into one of two categories, depending upon the specific attribute of interest. For instance, to measure attitudes toward flowers and insects, participants were asked to categorize the names of flowers and insects into the categories ‘flower’ or ‘insect’; and to measure attitudes toward race, participants were asked to categorize people’s names into typical ‘Black’ or ‘White’ names.

Importantly, participants categorized the stimuli in the diagnostic task by pressing the same response keys as those that were used to indicate whether words were ‘pleasant’ or ‘unpleasant’ in the inducer task. In one condition, one category (e.g. ‘flowers’) shared a response key with pleasant words, and the other category (‘insects’) shared the other response key with unpleasant words. In the other condition, the assignment of categories to response keys was reversed, such that flowers now shared a response key with unpleasant words and insects with pleasant words. The results showed that responses were faster with the former pairing of

the inducer categories and the diagnostic categories than with the latter pairing. This advantage of the former pairing was interpreted as reflecting the consistency of category assignment with implicit attitudes of the participants toward flowers and insects. That is, the flower-pleasant/insect-unpleasant assignment was *compatible* with implicit attitudes that participants hold toward flowers and insects, whereas the flower-unpleasant/insect-pleasant assignment was *incompatible* with them.

How implicit are implicit attitudes?

The IAT effect is usually attributed to two types of cognitive processes that compete for a response, namely automatic and controlled processes (e.g., De Houwer et al., 2009; Devine, 1989; Greenwald et al., 1998; Sherman et al., 2008). As is assumed for more widely known interference effects in cognitive psychology, such as Stroop interference (MacLeod, 1991) and the Simon effect (Lu & Proctor, 1995; Yamaguchi & Proctor, 2012), automatic processes are initiated with little attentional resources or intention regardless of the task context. Controlled processes are those that require attentional resources and depend on the intention to carry out a specific task. In the IAT, the controlled process is thought to operate based on the assignment of stimuli to response keys (e.g., 'flowers' to the left key and 'insects' to the right key), whereas the automatic process activates a response based on implicit associations between the stimuli and the concepts attached to the response keys (e.g., 'flowers are good' and 'insects are bad'). With the assignment of categories that is compatible with implicit associations, both controlled and automatic processes activate the correct response; with the assignment that is incompatible with implicit associations, the controlled process activates the correct response while the automatic process activates an incorrect response. As the automatic process reacts quickly, the tendency to

follow the automatic reaction has to be overcome if one is to make the correct response (Conray, Sherman, Gawronski, Hugenberg, & Groom, 2005).

Although it is a very popular and common notion within psychology, the term 'automaticity' does require careful elaboration. Traditionally, automatic processes are those that do not require attention, awareness, or intention (e.g., Kornblum, Hasbroucq, & Osman, 1990; MacLeod, 1991; Posner & Snyder, 1975; Schneider & Shiffrin, 1977), but no phenomenon has been found to satisfy *all* of these qualities of automaticity simultaneously (Bargh, 1989; Kahneman & Treisman, 1984; Logan, 1988; Melnikoff & Bargh, 2018; Moors & De Houwer, 2006). In the context of the IAT and related studies, the implicitness of 'implicit attitude' has been questioned by many researchers (De Houwer et al., 2009; Fazio & Olson, 2003; Fiedler, Messner, & Bluemke, 2006; Hahn, Judd, Hirsh, & Blair, 2014; Livingston & Brewer, 2002; Olson & Fazio, 2003). As a representative example, Fazio and Olson (2003) have shown that requiring categorization of faces in terms of race prior to the IAT inflates the IAT score, presumably due to increased attention to racial attributes of individuals. They argued that what is implicit in the IAT is not the attitude that the IAT aims to measure, but it is the fact that individuals' attitudes are being assessed; that is, participants who perform the IAT may be unaware that their attitudes are measured in the task, although they may be well aware of their attitudes toward the attributes used in the test (Berger, 2018; Hahn et al., 2014). Also, when the task-context is related to a specific attribute, the motivation to control a negative attitude toward that attribute is evoked, which overcompensates the negative bias (Wegener & Petty, 1995). The requirement to explicitly categorize materials based on a specific attribute may indeed evoke their attitudes toward that attribute in the IAT (Fazio, 1989; Klauer, 1997).

Influences of Explicit Categorization in the IAT

It is even possible that the IAT score depends *entirely* on the requirement to attend explicitly to a specific attribute of materials. Consistent with this possibility, no IAT effect was obtained based on the pleasantness of materials when the task required explicit categorization of the names of historical figures into British or non-British (De Houwer, 2001). Similarly, when the same set of materials were categorized based on two different attributes ('Black' vs. 'White', or 'athlete' vs. 'politician'), the IAT effect depended on the attribute that the task required to categorize the materials (Mitchell, Nosek, & Banaji, 2003). The IAT effect could even be 'reversed' if a positive category (e.g., 'flower') consisted of negative exemplars ('poison ivy') and a negative category ('insect') consisted of positive exemplars ('butterfly'; Govan & Williams, 2004). Similar results were obtained for the attribute materials (i.e., when positive inducer words are related to a negative category; Bluemke & Friese, 2006; Steffens & Plewe, 2001). Hence, the IAT effect depends on how the target and attribute categories are conceptualized, rather than the attitude toward the categories being implicit or automatic. Therefore, there appear to be abundant evidence that the *explicit* categorization required in the IAT can have a considerable impact on the IAT scores.

These findings suggest an important possibility that could undermine the validity of the IAT as a measure of implicit attitude to be manifested in everyday decision making. If the IAT scores depend on explicit categorization that is required by the task itself, then the IAT score may not be a valid representation of one's implicit attitude that is expressed in a situation where the attribute associated with the implicit attitude is irrelevant to making a judgment. For example, people who obtain a high score in the racial IAT may express a racial bias only in a condition where they are required to judge individuals based on race, but they *may* not suffer from the racial bias in a situation where they are required to judge the same individuals based on

another quality (e.g., selecting job candidates based on academic qualifications). The present study addresses this issue experimentally and asks whether the IAT effect can occur based on attributes of the stimuli that are irrelevant to the task at hand.

The Present Study

The aims of the present study were twofold; (1) whether the IAT effect is obtained based on a task-irrelevant attribute of the materials, and (2) how the explicit requirement to categorize the materials based on one attribute influences the IAT effect based on a task-irrelevant attribute. To this end, the present study used a multi-attributes version of the IAT task, or *m-IAT* (also see De Houwer, 2001; Mitchell et al., 2003). As in the original version of the IAT, the m-IAT requires participants to perform two tasks, the inducer task and the diagnostic task. The inducer task associates non-evaluative responses (left and right key presses) with evaluative qualities ('pleasant' vs. 'unpleasant'); participants have to categorize a set of words into 'pleasant' or 'unpleasant' categories according to their meanings by pressing two alternative keys. The diagnostic task tests the pre-existing associations between social or non-social attributes of stimuli and the evaluative qualities of the responses; participants categorize images of individuals' faces according to an attribute by pressing the response keys that are also used in the inducer task. More specifically, in the present study, participants categorized photographs of individuals based on their gender or race, and we examined whether the IAT effect emerged on another attribute that was irrelevant to the categorization. This version of the IAT provided a means to address not only the issue of whether explicit categorization of an attribute is necessary to obtain the IAT effect on that attribute, but also the issue of how explicit categorization of one attribute influences the IAT effect on other attributes.

Previous studies that used the IAT task involving multiple attributes found that the IAT effect emerged only on an attribute that was relevant to the task (De Houwer, 2001; Mitchell et al., 2003). These are indeed curious results, given that there are many variations of cognitive tasks that yield effects of task-irrelevant attributes of stimuli on performance (e.g., Kohnblum et al., 1990; MacLeod, 1991; Lu & Proctor, 1995), including those that are concerned with implicit attitudes, such as the affective misattribution procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005), the affective priming task (Fazio, Sanbonmatsu, Powell, & Kardes, 1986), the Go/No-Go association test (Nosek & Banaji, 2001), the extrinsic affective Simon task (De Houwer, 2003), the pronunciation task (Bargh, Chaiken, Raymond, & Hymes, 1996), and others (see Nosek et al., 2011). One possibility is that this is due to the unique procedure used in the IAT. Unlike the IAT, most other procedures avoid explicit categorization or evaluation of the target attribute in the task (see Livingston & Brewer, 2002), and this requirement of explicit categorization might eliminate the IAT effect based on task-irrelevant attributes.

Evidence supporting such a possibility is provided in a classic study (Hedge & Marsh, 1975), in which an advantage of task-irrelevant stimulus-response compatibility (i.e., the Simon effect) was reversed by task instructions. In that study, participants were required to respond to the colors of stimuli (red and green lights) by pressing keys that were colored in the same as, or different from, the stimuli. When participants pressed the key of the same color as the stimulus (i.e., red response to red light, and green response to green light), there was a typical spatial compatibility effect; that is, responses were faster to spatially-compatible stimuli than to spatially-incompatible stimuli, even when the spatial attribute was task-irrelevant. However, when participants pressed the keys in the alternate colors, the spatial compatibility effect was reversed to favour paring of incompatible stimulus and response locations (i.e., left key press

was faster for stimuli on the right than on the left). This reversal suggests that the spatial compatibility effect depended on explicit color categorization that participants intended to perform (see also Baroni, Yamaguchi, Chen, & Proctor, 2013). Similar influences of explicit categorization might have been present in previous studies of the IAT that also involved task-irrelevant attributes, but the task-irrelevant IAT score might be averaged out when the results for the compatible and incompatible blocks were aggregated (Mitchell et al., 2003)[1]. Therefore, we consider whether the explicit categorization required in the task influences the IAT effect based on attributes of stimuli that are irrelevant to the required categorization.

Across four experiments reported in the present article, the attributes used in the diagnostic task were either social (race or gender) or non-social (color) attributes of stimuli. Experiment 1 aimed to provide the basic outcomes of the m-IAT with regard to the IAT effects based on the task-relevant and task-irrelevant social attributes. The results of Experiment 1 would also be suggestive of how the explicit requirement to categorize materials based on one attribute influences 'implicit' categorizations based on another attribute (e.g., how categorization based on race influences the IAT effect based on gender bias). Experiment 2 further tested whether 'implicit' categorization of materials based on a task-irrelevant attribute would modulate 'implicit' categorization based on another task-irrelevant attribute. Experiments 3 and 4 followed up the findings of these two experiments and examined the influence of explicit categorization of non-social attributes on 'implicit' categorization of social attributes. These experiments reveal important clues about the underlying mechanisms of the IAT.

[1] De Houwer (2001) also examined whether the IAT effect emerged based on a task-relevant attribute or a task-irrelevant attribute. He used two dichotomous values (British vs. non-British, and like vs. dislike) and found the effect based only on the task-relevant attribute, with no sign of the effect on the task-irrelevant attribute. However, this study did not test a condition in which the two attributes were flipped in their roles, so it does not exclude the possibility that the IAT effect might not have emerged even when the task-irrelevant attribute was used as the task-relevant attribute.

Experiment 1

In Experiment 1, stimuli for the inducer task were a set of words, which were categorized into 'pleasant' or 'unpleasant' based on their meaning. Stimuli for the diagnostic task were images of individuals that varied in both race and gender (White females and males, and Black females and males). Each participant performed two versions of the diagnostic task. In the *race task*, the materials were categorized in terms of the racial groups (White vs. Black); in the *gender task*, the materials were categorized in terms of the gender groups (male vs. female). Only participants who identified their own race as White were included in the present experiment (32 males and 32 females).

From previous IAT studies (e.g., Greenwald et al., 1998), it was expected that responses should be faster when participants pressed the 'pleasant' and 'unpleasant' keys, respectively, for individuals who belong to their own racial or gender group and for those who do not belong to their own racial or gender group than when they pressed the same keys with the reversed assignments. That is, as all participants were identified themselves as White, they should press the 'pleasant' key for White individuals and the 'unpleasant' key for Black individuals faster than the 'unpleasant' key for White individuals and the 'pleasant' key for Black individuals; this is the *race IAT effect*. Also, participants should press the 'pleasant' key to their own gender group and the 'unpleasant' key to the opposite gender group faster than the 'unpleasant' key to their own gender group and the 'pleasant' key to the opposite gender; this is the *gender IAT effect*. Several studies have shown that the gender IAT effect is typically found amongst females but not males (Aidman & Carroll, 2003; Meissner & Rothermund, 2013; Mitchell et al., 2003; Nosek & Banaji, 2001). Although the reason for such findings is still open to debate, the results of female and male participants were considered separately in the present experiment.

Just as in the Simon task (Hedge & Marsh, 1975), the IAT effect may also depend on whether category assignments are compatible or incompatible with participants' attitude toward the target attribute. In the race IAT task, the task-irrelevant gender attribute may produce a standard IAT effect (i.e., faster responses when participants' own gender group is responded to by pressing the 'pleasant' key than when it is responded to by pressing the 'unpleasant' key) in a block where the task-relevant race attribute is mapped to the pleasant/unpleasant categories compatibly (i.e., 'White' is assigned to 'pleasant' and 'Black' is assigned to 'unpleasant'). But the task-irrelevant gender attribute may produce a reversed IAT effect (i.e., faster responses when participants' own gender group is responded to by pressing the 'unpleasant' key than when it is responded to by pressing the 'pleasant' key) in a block where the task-relevant race attribute is mapped to the pleasant/unpleasant categories incompatibly ('White' is assigned to 'unpleasant' and 'Black' is assigned to 'pleasant'). If a reversal of the task-irrelevant IAT effect occurs in the incompatible block, it could have cancelled out the effect in the compatible block when these blocks were aggregated together. This might explain the lack of the IAT effect based on a task-irrelevant attribute in previous studies (e.g., De Houwer, 2001; Mitchell et al., 2003).

Method

Participants

Sixty four participants were recruited for the present experiment[2]. Half of the participants were White males (mean age = 21.72, *SD* = 2.94, range = 18-30), and the other half were White females (mean age = 20.47, *SD* = 2.72, range = 18-32). Their racial identity was based on self-report. All participants were students at Edge Hill University in the Northwest of

[2] With all variables being within-subject factors, a sample size of 24 gives a power of greater than .99 for a large effect size and .95 for a medium effect size, assuming a correlation of .8 among repeated measures. To be conservative, we decided to recruit a larger sample size of 32 per group in the first experiment and a sample size of more than 24 in each of the subsequent experiments.

England and were either paid £3 or received experimental credits toward their psychology module for participation. All participants reported having normal or corrected-to-normal visual acuity and color vision. They were naïve as to the purpose of the experiment. The experimental protocol was approved by the Research Ethics Committee of the Department of Psychology at Edge Hill University. In this and subsequent experiments, we report all measures, manipulations, and exclusions. Sample size was determined before any data analysis.

Apparatus and Stimuli

The apparatus consisted of a personal computer and a 23-in. widescreen monitor. The experiment was controlled by E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). Stimuli were photographs of 160 individuals (40 White females, 40 White males, 40 Black females, and 40 Black males), selected from the *10K US Adult Faces Database* (Bainbridge, Isola, & Oliva, 2013). All faces expressed happy or neutral emotions. These images were matched on five key attribute ratings (see Table 1). Images were adjusted to fit within 256 x 256 pixels and were displayed at the center of screen. There were five positives words (happy, cheer, peace, love, and pleasure) and five negatives words (filth, evil, murder, abuse, and agony). They were also presented at the center of screen, in the Courier New font at 18-pt. font size and printed in black against a white background. The category labels were presented at the top left and right corners of the screen, indicating the assigned responses (‘z’ or ‘/’ keys). The labels were ‘Good’ and ‘Bad’ for the inducer task, ‘Black’ and ‘White’ for the race task, and ‘Female’ and ‘Male’ for the gender task. When the inducer task was mixed with either of the latter tasks, the labels of both tasks were presented simultaneously; the labels for the inducer task were always presented above the labels of the diagnostic task. They were also in the 18-pt. Courier New font, printed in black.

Responses were registered by pressing the ‘z’ and ‘/’ keys on a standard QWERTY keyboard with the left and right index fingers, respectively.

Procedure

The experiment was conducted individually in a cubicle under normal fluorescent lighting. Participants sat in front of the computer monitor at an unrestricted distance of 50 cm and read on-screen instructions. There were two phases in a session. One phase involved the race task in which participants categorized the race of the face stimuli into ‘White’ or ‘Black’; the other phase involved the gender task in which participants categorized the gender of the face stimuli into ‘male’ or ‘female’. These tasks were designated as the diagnostic tasks. Each phase also involved the inducer task.

Each phase started with a practice block of eight trials on the inducer task. Half the trials presented positive words, and the other half presented negative words, in a random order. The locations of category labels (‘Good’ and ‘Bad’) were counterbalanced across participants. The next block consisted of eight practice blocks of the diagnostic task (race or gender), with an equal number of images of White males, White females, Black males, and Black females. The category labels were ‘Black’ and ‘White’ or ‘Male’ and ‘Female’, respectively, with the locations being counterbalanced across participants as well. The third block mixed the inducer task and the diagnostic task in a random order. There were 12 trials in this block, four trials for the inducer task and eight trials for the diagnostic task. The fourth block was the same condition as the third block, consisting of 90 test trials (30 trials for the inducer task and 60 trials for the diagnostic task). The fifth block was also the diagnostic task, but the locations of the category labels were switched; this block also consisted of eight practice trials. The sixth block consisted of 12 practice trials of the inducer and diagnostic tasks with the switched category labels for the

diagnostic task; the category labels for the inducer task remained the same throughout the phase. The final block was the same as the sixth block, except that there were 90 test trials (30 trials for the inducer task and 60 trials for the diagnostic task). The second phase was essentially the same as the first phase, except that the diagnostic task was replaced with the other task (race or gender). The order of the race and gender tasks was counterbalanced across participants.

Each trial started with a fixation cross at the center of screen for 750 ms, followed by the imperative stimulus with the category labels at the top left and right corners. The stimulus remained on the screen for 3000 ms or until a response was made. If the response was incorrect, the error message "Error!" was presented for 500 ms; if there was no response, the message "Faster" was presented; and if the response was correct, a blank screen appeared for 500 ms. RT was the interval between onset of the stimulus and a depression of a response key. Response accuracy was also recorded on each trial.

Results

The analysis focused on trials of the diagnostic tasks. Trials for which RT was less than 200 ms or longer than 2000 ms were discarded from the analysis (0.74% of all trials). The overall error rate was 4.99%.

For both diagnostic tasks (gender and race), mean RT for correct responses and percentage of error trials (PE) were computed for each participant in terms of the compatibility between the race category and the inducer category (Race Compatibility: compatible vs. incompatible) and the compatibility between the gender category and the inducer category (Gender Compatibility: compatible vs. incompatible). For Race Compatibility, as all participants were White, trials were compatible if participants pressed the key assigned to 'Good' for 'White' or the key assigned to 'Bad' for 'Black'; trials were incompatible if they pressed the key

assigned to ‘Good’ for ‘Black’ or key assigned to ‘Bad’ for ‘White’. Gender Compatibility depended on the gender of the participants. Trials were compatible if participants pressed the key assigned to ‘Good’ for own gender or the key assigned to ‘Bad’ for the opposite gender; trials were incompatible if they pressed the key assigned to ‘Good’ for the opposite gender or the key assigned to ‘Bad’ for their own gender. RTs were submitted to a 2 (Task: race vs. gender) x 2 (Race Compatibility) x 2 (Gender Compatibility) ANOVA. All factors were within-subject variables. Note that several previous studies found the gender IAT effect for female participants but not for male participants (e.g., Meissner & Rothermund, 2013; Mitchell et al., 2003), so we analyzed the data separately for female and male participants. RT and PE are summarized in Figure 1 and in Table 2, respectively. The results of ANOVAs are summarized in Table 3.

Female participants.

RT depended on Gender Compatibility and Race Compatibility, but both compatibility effects were modulated by Task. For the gender task, the task-relevant gender IAT effect was 76 ms, and the task-irrelevant race IAT effect was 7 ms. For the race task, the task-relevant race IAT effect was 55 ms, and the task-irrelevant gender IAT effect was 21 ms. Thus, the IAT effects were larger when the corresponding dimension was task-relevant than when it was task-irrelevant in general. The two types of compatibility also interacted. Although these two factors did not interact with Task, interesting patterns did emerge when the interaction between the two types of compatibility is assessed separately for the two tasks (see Figure 1). For the gender task, the interaction reflected the outcome that there was 40 ms of the task-irrelevant race IAT effect ($p < .001$) when the genders were mapped compatibly with the inducer category, but it was reversed to –27 ms ($p = .013$) when the genders were mapped incompatibly with the inducer category. Similarly, for the race task, there was 51 ms of the task-irrelevant gender IAT effect (*p*

< .001) when the races were mapped compatibly with the inducer category, but it was reduced to –9 ms (p = .463) when the races were mapped incompatibly with the inducer category. In short, the task-irrelevant IAT effect was significant for the blocks with compatible mappings but was reversed or eliminated for the blocks with incompatible mappings (for which participants had to overcome the implicit bias toward the task-relevant categories).

PE showed that responses were generally more accurate for the gender task (M = 4.62%) than for the race task (M = 7.44%). Responses were also more accurate when gender was compatible with the inducer category (M = 5.29%) than when it was incompatible (M = 6.76%), but there was a non-significant trend that this gender IAT effect depended on the task; there was no effect (M = 0%) for the gender task and 2.5% for the race task. No other effect reached significance.

Male participants.

Consistent with a number of previous studies (e.g., Aidman & Carroll, 2003; Mitchell et al., 2003; Nosek & Banaji, 2001), Gender Compatibility produced only a marginal effect, whereas Race Compatibility produced a significant effect. However, both compatibility effects depended on Task (although the interaction was only marginal for Gender Compatibility). The gender IAT effect was 39 ms (p = .047) for the gender task and it was –2 ms (p = .697) for the race task; the race IAT effect was 101 ms (p < .001) for the race task and it was 14 ms (p = .054) for the gender task. No other effects were significant. As was the case for female participants, the task-irrelevant IAT effect was examined separately for the compatible and incompatible mapping blocks. For the gender task, the task-irrelevant race effect was 28 ms (p = .006) in the compatible mapping block and 0 ms (p = .991) in the incompatible mapping block. For the race task, the task-irrelevant gender effect was 7 ms (p = .352) in the compatible mapping block and -

3 ms (p = .721) in the incompatible block. Thus, the race IAT effect still occurred even when the race was an irrelevant attribute, and the effect depended on whether the task-relevant gender attribute was mapped compatibly or incompatibly to the inducer category. The gender IAT effect did not occur reliably for male participants regardless of their relevance to the task.

PE indicated that responses were more accurate for the gender task (M = 3.21%) than for the race task (M = 4.70%), and when race was compatible with the inducer category (M = 3.29%) than when it was incompatible (M = 4.62%), yielding the race IAT effect. Although the gender IAT effect was not significant overall, it depended on the task; the gender IAT effect was 1% for the gender task and it reversed to -1.40% for the race task. There was also a larger race IAT effect when gender was compatible with the inducer category (M = 2.17%) than when it was incompatible (M = .05%). No other effects were significant.

Discussion

The present experiment demonstrated typical gender and racial IAT effects based on the task-relevant attribute in RT. Thus, responses were generally faster when the race or gender to which the participants belonged was mapped to the category 'pleasant' than when it was mapped to the category 'unpleasant.' The experiment also showed that IAT effects emerged on task-irrelevant attributes. When participants categorized images based on the gender of the relevant stimulus, images were categorized to 'pleasant' faster when the race of the individuals in the images was the same as participants' own race than when it was different. Furthermore, when participant categorized images based on the race of the relevant stimulus, images were categorized to 'pleasant' faster when the gender of the individuals in the images was the same as the participants' own gender than when it was different. Therefore, task-irrelevant attributes of stimuli still influence how quickly participants categorized images, producing task-irrelevant

IAT effects. Note that the effects of task-irrelevant attributes were generally smaller than those of task-relevant attributes. The results are thus consistent with previous findings that the IAT effect depends on how stimuli are categorized (Mitchell et al., 2003), but this study extends these findings by showing that the IAT effect still occurs even when participants are not required to attend to the attribute that produced the effect. However, it is unclear if attention is required for the task-irrelevant IAT effect (this is examined subsequently in Experiment 4).

The present experiment also indicated that these task-irrelevant IAT effects depended on the compatibility of the explicit categorization. In the block for which categorization was compatible with biases (White-pleasant/Black-unpleasant or own gender-pleasant/opposite gender-unpleasant), task-irrelevant IAT effects were also consistent with the biases; White and Black were responded faster, respectively, with pleasant and unpleasant responses than the reverse, whereas own gender and opposite gender were responded faster with pleasant and unpleasant responses than the reverse. In another block for which categorization was incompatible with biases (White-unpleasant/Black-pleasant or own gender-unpleasant/opposite gender-pleasant), the task-irrelevant IAT effects were reversed; White and Black were now responded faster, respectively, with unpleasant and pleasant responses than vice versa, whereas own gender and opposite gender were responded faster with unpleasant and pleasant responses than vice versa. Hence, task-irrelevant IAT effects depended on whether explicit categorization was compatible or incompatible with biases. These outcomes are consistent with the classic spatial compatibility effect (Simon effect), which also depends on explicit categorization of the task-relevant stimulus features (Hedge & Marsh, 1975). This finding is important because it raises the possibility that the IAT effect can be modified in the way that non-social bias is controlled (e.g., Yamaguchi, Chen, & Proctor, 2015).

The reason for the reversal requires further scrutiny. It has been suggested that multiple cognitive processes are involved in performing the IAT - detection of the task-relevant attribute, activation of stereotypical associations, general response bias (or guessing), and the overcoming of any activated stereotype (Sherman et al., 2008). In a block with the incompatible category assignment (e.g., 'Black-pleasant'/'White-unpleasant'), participants have to suppress activated associations to overcome an automatic response tendency. Such efforts may modulate the influence of a task-irrelevant attribute as well, resulting in the observed reversal of the IAT effect. It is also possible that the reversal resulted from the consistency between the task-relevant and task-irrelevant implicit associations[3]. In the compatible block, responses compatible with the task-irrelevant attribute are also compatible with responses required by the task-relevant attribute; thus, the task-relevant and task-irrelevant associations are consistent (i.e., both compatible). But responses incompatible with the task-irrelevant attribute are still compatible with responses required by the task-relevant attribute; thus, the task-relevant and task-irrelevant associations are inconsistent (one is compatible but the other is incompatible). In the incompatible block, responses compatible with the task-irrelevant attribute are now incompatible with responses required by the task-relevant attribute, so the task-relevant and task-irrelevant associations are consistent. But responses incompatible with the task-irrelevant attribute are compatible with responses required by the task-relevant attribute, so the task-relevant and task-irrelevant associations are inconsistent. If responses are faster when the task-relevant and task-irrelevant associations are consistent than when they are inconsistent, this consistency between the task-relevant and task-irrelevant implicit associations could also explain the reversed IAT effect. This possibility is considered in Experiment 2.

[3] We thank Jan De Houwer for pointing this out.

Moreover, it should be noted that, with the design of the present experiment, there were potential confounding effects of stimulus attributes and the task-irrelevant IAT effect particularly for female participants[4]. In the race task, the task-irrelevant (gender) IAT effect was computed by subtracting RT for Black males and White females from RT for Black females and White males in the compatible block, and by subtracting RT for Black females and White males from RT for Black males and White females in the incompatible block. Similarly, in the gender task, the task-irrelevant (race) task was computed by subtracting RT for Black males and White females from White males and Black females in the compatible block, and by subtracting RT for Black females and White males from RT for White females and Black males in the incompatible block. Consequently, the task-irrelevant IAT effect could be positive in the compatible block and negative in the incompatible block, just because RT for Black females and White males were longer in general than RT for black males and white females for unknown stimulus characteristics. Note, however, this applies exclusively to the results of female participants, and we should have obtained a negative task-irrelevant IAT effect in the compatible block and a positive task-irrelevant IAT effect in the incompatible block for male participants, if such confounding factors were responsible for the task-irrelevant IAT effects. The present results did not indicate such biases. This issue is also considered in Experiment 2 further.

To summarize, Experiment 1 revealed that (1) social biases do influence people's judgement even when the attributes are irrelevant to the task at hand, but the effects of task-irrelevant attributes are much smaller than those of task-relevant attributes, and that (2) the direction of the effect of an irrelevant social bias depends on whether the task requires participants to overcome the social bias against the task-relevant attribute; when participants

[4] We thank Christoph Klauer for pointing this out.

categorize the individuals against their attitudes toward the individuals' gender or race, the task-irrelevant IAT effect is also reversed. These outcomes suggest that participants' intention to counter negative biases toward one attribute can also reverse the influences of negative biases toward another attribute that is present simultaneously even if they are not focal characteristics of the task.

Experiment 2

As shown in Experiment 1, the IAT effect occurs based on a task-irrelevant attribute even when participants were not required to use the attribute to categorize individuals. This implies that participants categorized the individuals implicitly based on an irrelevant attribute. Interestingly, the results of Experiment 1 also suggested that this implicit categorization depended on explicit categorization based on the task-relevant attribute, such that the task-irrelevant IAT effect could be reversed, or eliminated, when explicit categorization required to overcome implicit attitude toward the task-relevant attribute. Experiment 2 addressed the question of whether the task-irrelevant IAT effect also depended on 'implicit' categorization of another task-irrelevant attribute.

The IAT effect based on a task-irrelevant attribute could have been reversed for two possible reasons. The first possibility is that it depends on one's intention to overcome implicit attitude toward the task-relevant attribute. When categorizing a Black male into the 'pleasant' category based on race, for example, categorization of the race ('Black' to 'pleasant') requires an explicit intention to overcome a racial bias. Although the task-irrelevant gender IAT effect implies that implicit categorization of the gender ('male' to 'pleasant') did occur, it may not require an explicit intention to overcome a gender bias. Therefore, if an intention to overcome a bias is required to eradicate or reverse the effects of another social bias, it is unlikely that the

IAT effect is modulated on trials that require incompatible response to a task-irrelevant attribute. However, if an intention is not required to reverse a task-irrelevant IAT effect, the IAT effect should depend also on whether a task-irrelevant attribute is responded compatibly or incompatibly.

The second possibility is that the reversal of the IAT effect obtained in Experiment 1 might have been because of the consistency of two implicit associations; that is, responses are faster when they are incompatible to both the task-relevant and -irrelevant attribute than when they are compatible with one attribute but incompatible with the other. This could also explain the patterns of the results in Experiment 1. If this is the case, the consistency of two task-irrelevant implicit associations should also result in a reversed IAT effect.

The present experiment extended the m-IAT in Experiment 1 and provided a test of these two possibilities. The task included one task-relevant attribute and two task-irrelevant attributes. As a pilot study, we tested a variation of the m-IAT using the color (green/red, along with other color pairs such as purple/pink and blue/yellow) as the task-relevant attribute and found that categorization was consistently faster when green and red were categorized, respectively, into 'pleasant' and 'unpleasant' than when green and red were categorized into 'unpleasant' and 'pleasant.' There are several possible reasons for the advantage of the former assignment: for instance, green is often used as a 'go' signal whereas red is used as a 'stop' signal; green is often used to indicate 'correct' whereas red is used to indicate 'incorrect'; or there is the possibility that the green we used in the pilot experiment was inadvertently brighter than the red. Regardless of the reason for the associations, it is sufficient for our present purposes that there are consistent biases in categorizing these particular colors into pleasant and unpleasant responses.

In the present experiment, stimulus images appeared within a colored frame that changed into green or red across trials. Participants were all female because males provided less clear results in Experiment 1 with regard to the effect of gender (see also Aidman & Carroll, 2003; Meissner & Rothermund, 2013; Mitchell et al., 2003; Nosek & Banaji, 2001). The task was essentially the same as Experiment 1; one phase required explicit categorization of images based on race, and the other phase required explicit categorization of the same images based on the color of the frames. We assessed (1) whether more than one task-irrelevant attribute could influence categorization performance simultaneously (i.e., whether there could be task-irrelevant IAT effects based on multiple attributes), (2) whether explicit categorization affects more than one task-irrelevant attribute simultaneously, and (3) whether compatibility of a task-irrelevant attribute and categorization response influences the IAT effect based on another task-irrelevant attribute. Note that green and red frames were used for all images with an equal probability; thus, there were no confounding effects of color compatibility and gender/race of individuals, which might have been present as in Experiment 1 (see the Discussion of Experiment 1). The occurrence of a task-irrelevant IAT effect based on the frame color and its reversal in the compatible and incompatible blocks of the race task in the present experiment would indicate that the task-irrelevant IAT effect does depend on compatibility of a task-irrelevant attribute with the response category, not on unknown characteristics of the stimuli facilitating response speed for Black females and White males. Finally, it is noteworthy that there is a possibility that task-irrelevant IAT effects occurred in Experiment 1 because task-irrelevant attributes in one phase was used as the task-relevant attribute in the other phase. Given that gender was never task-relevant in the present experiment, such a possibility could be excluded if task-irrelevant gender IAT effect is to be replicated in the present experiment.

Method

Participants

Twenty-five white females (mean age = 21.28, *SD* = 3.29, range = 18-31) were recruited from the Edge Hill University community. They were either paid £3 or received experimental credits toward their psychology module for participation. All participants reported having normal or corrected-to-normal visual acuity and color vision, and none had participated in Experiment 1.

Apparatus, Stimuli, and Procedure

The experiment was similar to that of Experiment 1, but the stimulus display consisted of an image of a Black or White individual (as used in Experiment 1) with an additional color frame (green or red) surrounding the image. The photographs of individuals were the same as those in Experiment 1, and these photographs appeared in a green or red frame with an equal probability. The gender task was replaced with the color task, in which participants categorized the color frame into the 'Green' or 'Red' categories. Thus, the gender was always task-irrelevant in both tasks. The procedure followed that of Experiment 1 closely in other respects.

Results

The data were filtered (0.34% of all trials) and analyzed in the same manner as in Experiment 1. The overall error rate was 6.27%. Mean RTs for correct responses and PEs were submitted to 2 (Race Compatibility) x 2 (Gender Compatibility) x 2 (Color Compatibility) ANOVAs separately for the race and color tasks. Based on the pilot experiment and a preliminary inspection of the current data, color compatibility was coded as compatible for green-pleasant/red-unpleasant mappings and as incompatible for green-unpleasant/red-pleasant

mappings, which showed consistent compatibility relationships. RT and PE are summarized in Figure 2 and Table 4. The results of ANOVA are summarized in Table 5.

Race task

RT indicated that for the race task, there was a significant race IAT effect, yielding a 60-ms advantage for the White-pleasant/Black-unpleasant mapping (M = 637 ms) compared to the White-unpleasant/Black-pleasant mapping (M = 697 ms). This effect represented the task-relevant IAT effect. Although there were no main effects of Gender Compatibility or Color Compatibility, these factors interacted with Race Compatibility, indicating the modulations of the race and color IAT effects according to the category assignment for race categorization. The color IAT effect was 24 ms when the mapping between the race and inducer categories were compatible, but it reversed to –29 ms when the mapping was incompatible. Similarly, the gender IAT effect was 39 ms when the mapping between the race and inducer categories were compatible, but it reversed to –25 ms. These outcomes represented the modulations of task-irrelevant IAT effects based on the compatibility of task-relevant category assignment, which was found in Experiment 1, and the findings of the color IAT effect indicate that these patterns do not reflect differential response speeds due to unknown stimulus properties as discussed in Experiment 1. Interestingly, although the effects of color and gender were obtained in the present experiment, there was no interaction between these two task-irrelevant attributes. Thus, the task-irrelevant IAT effects depended on categorization of the task-relevant attribute but not categorization of the task-irrelevant attribute. PE showed no significant effect.

Color task

For RT, the only significant effect was a main effect of Color Compatibility. RT was 98 ms shorter for the green-pleasant/red-unpleasant mapping (M = 519 ms) than for the green-

unpleasant/red-pleasant mapping (*M* = 617 ms). The lack of the race and gender IAT effects in this task is interesting, as it suggests that there was no task-irrelevant IAT effects.

PE also showed a significant color IAT effect; responses were more accurate for the green-pleasant/red-unpleasant mapping (*M* = 4.70%) than for the green-unpleasant/red-pleasant mapping (*M* = 7.42%). There was also a significant task-irrelevant gender IAT effect, but its direction was opposite to what would be expected; responses were more accurate for gender incompatible trials (*M* = 5.44%) than for gender compatible trials (*M* = 6.68%). No other effects were significant.

Discussion

The results of this experiment demonstrated task-relevant IAT effects based on race and color. In the race task, the task-irrelevant IAT effects were also obtained based on gender and color, and the occurrence of the task-irrelevant color IAT effect ruled out the possible confounding effect of an unknown stimulus property that might have slowed down responding to White males and Black females, which could have explained the task-irrelevant IAT effects of Experiment 1. Thus, the present results reinforce the earlier conclusion that the IAT effect does occur based on task-irrelevant attributes. The results also corroborated the finding in Experiment 1 that the task-irrelevant IAT effects depended on explicit categorization of the task-relevant attribute, such that the task-irrelevant IAT effects were reversed when the task required participants to counter implicit attitude. In addition, as gender was never task-relevant in the present experiment, the task-irrelevant gender IAT effect could not depend on participants' prior experiences to perform the gender task. Therefore, the present results confirm that more than one task-irrelevant attribute can influence categorization simultaneously, and the influences of task-irrelevant attributes depend on explicit categorization of the task-relevant attribute.

Furthermore, Experiment 2 also showed that the task-irrelevant IAT effects did not depend on the compatibility of the other task-irrelevant attribute; that is, categorizing individuals against the task-relevant attribute reversed the IAT effect based on a task-irrelevant attribute, but categorizing the same individuals against a task-irrelevant attribute did not reverse the IAT effect based on another task-irrelevant attribute. This is a curious finding because the occurrence of the task-irrelevant IAT effect implies that participants categorized stimuli based on the task-irrelevant attribute and would need to overcome a bias if the response is incompatible with the bias. The results indicate that it is not the consistency of two implicit associations that are responsible for the results but agree with the claim that it is explicit categorization of the task-relevant attribute, or the intention to overcome the bias toward the attribute, that modulates task-irrelevant IAT effects.

In addition to these main findings of interest, the present experiment provided another interesting outcome that in the case of the color task, the only significant effect was that of the task-relevant attribute (i.e., color IAT effect) in RT. Neither of the two task-irrelevant attributes produced an IAT effect. Although the task-irrelevant gender yielded a significant effect in PE, its direction was opposite to what would be expected from the results of Experiment 1. The results suggest that the task-irrelevant IAT effect is somehow limited (also see De Houwer, 2001). There are several possible reasons for the lack of task-irrelevant IAT effects in the color task, and we considered three that we thought were equally plausible.

The first possibility that we considered was that color categorization is so fast that participants had no time to process the race or gender of the individuals in the images. In fact, the time that it took to categorize color ranged from 500 ms to 650 ms on average, whereas the time that it took to categorize images based on race ranged from 600 ms to 750 ms. This 100-ms

advantage of the color categorization might have eliminated the time for other attributes to influence responding (*relative speed hypothesis*). The second possibility was that social biases (i.e., racial or gender biases) influence categorization performance only when categorization is performed on characteristics of individuals that are related closely to social preferences (e.g., Bluemke & Friese, 2006; Govan & Williams, 2004; Spruyt, De Houwer, & Hermans, 2009). That is, the task-irrelevant IAT effect reflects *conditional automaticity* (see Bargh, 1989; Melnikoff & Bargh, 2018). It may be that the task-irrelevant gender attribute influences categorization based on race because both attributes are relevant to individuals' social characteristics, whereas the color of an image frame does not influence categorization of race or gender because it is not a social characteristic of the individuals (*conditional automaticity hypothesis*). The third possibility was that the influence of task-irrelevant attributes is simply a matter of perceptual salience. The more salient an attribute is, the more strongly the attribute influences categorization, regardless of its speed of processing or relatedness to the task-relevant attribute (*relative salience hypothesis*). Experiment 3 tested the relative speed hypothesis, and Experiment 4 tested the conditional automaticity hypothesis and the relative salience hypothesis.

Experiment 3

To test the relative speed hypothesis in the present experiment, we introduced variable delays between onsets of the color frames and the photographs of individuals in the color task used in Experiment 2. Each trial presented an image of an individual. With a delay of 100, 250, or 500 ms, the color frame appeared around the image. This delay provided a temporal advantage for the facial features of the individuals to be processed before the task-relevant color appeared. The delayed presentation technique has been used in the Stroop task to investigate the time course of information processing (Glaser & Glaser, 1982), which showed that there were

larger facilitating effects of compatible task-irrelevant word meanings when the words occurred before the task-relevant color. In the current task, if the relative speed hypothesis was correct, task-irrelevant social attributes (e.g., race and gender) would influence color categorization more profoundly and increases task-irrelevant IAT effects when the photographs were presented before the task-relevant color frames. If speed of processing does not explain the lack of the task-irrelevant IAT effects in the color task, then the temporal advantage should have little impact on the task-irrelevant IAT effects.

Method

Participants

Twenty-four White females (mean age = 24.71, *SD* = 5.39, range = 18-37) were newly recruited from the same subject pool, with the same selection criteria.

Apparatus, Stimuli, and Procedure

In the present experiment, the only diagnostic task was the color task. After the first three blocks of practice trials as in the previous experiments, there were three blocks of 108 test trials each (36 trials for the inducer task and 72 trials for the color task). For the color task, there were three types of trials with different color onset delays (CODs; i.e., the interval between the onset of the face and the onset of the color frame), which occurred randomly and equally often within each test block. The three CODs were 100, 250, and 500 ms.

In the color task, a trial started with the fixation cross for 750 ms, followed by a face image used in Experiments 1-2. With a variable COD, the color frame appeared around the image. Participants had 3,000 ms to make a response after the onset of the color. After the three test blocks, there were two practice blocks with the category labels for the color task being switched their locations, which were followed by another three test blocks. These test blocks

were essentially the same as the first three test blocks, except for the locations of the category labels for the color task.

Results

The data were filtered in the same manner as in the preceding experiments (0.53% of all trials). The overall error rate was 3.02%. Mean RTs and PEs were submitted to a 3 (Color Onset Delay: 100 ms, 250 ms, and 500 ms) x 2 (Race Compatibility) x 2 (Gender Compatibility) x 2 (Color Compatibility) ANOVA. Color compatibility was coded as in Experiment 2. RT and PE are shown in Figure 3 and in Table 6. The ANOVA results are summarized in Table 7.

The ANOVA revealed only two significant effects: main effects of Color Onset Delay and of Color Compatibility. All other effects were not significant. RT was faster with the green-pleasant/red-unpleasant assignment (M = 494 ms) than with the green-unpleasant/red-pleasant assignment (M = 589 ms). RT increased as the color onset delay decreased (Ms = 514, 535, and 575 ms, for 500-, 250-, and 100-ms delays, respectively). These outcomes indicate that face images interfered with responding to the color as the color onset was closer to the onset of the face images. Although this outcome implies that the faces interfered with processing color information at shorter delays, there was little evidence that the race or gender of the faces influenced responding to the color. There was no significant effect in PE.

Discussion

The results of the present study were clearly inconsistent with the relative speed hypothesis that the influences of task-irrelevant social attributes (race and gender) on explicit categorization of color was due to fast processing of color information. The present results indicated that the task-irrelevant IAT effects were absent even when there was sufficient time (> 100 ms) to compensate the slow processing speed of social attributes. It raises a question of

whether social attributes are actually processed in the color task for which only perceptual qualities are relevant to performing the task. It is important to note that the present finding also indicates that the racial IAT effect is not purely perceptually based (e.g., skin colors), given that it does not emerge when the color was the task-relevant attribute.

Experiment 4

In Experiment 4, we tested the two remaining hypotheses, the conditional automaticity hypothesis and the relative salience hypothesis. Although the traditional conception of automaticity is that certain stimuli are processed unintentionally or even against the intention not to do so (e.g., Posner & Snyder, 1975; Schneider & Shiffrin, 1977), such involuntary processes do not always occur unconditionally (Melnikoff & Bargh, 2018). Even in the Stroop and similar interference tasks that produce robust influences of task-irrelevant stimuli on performance, processing task-irrelevant stimulus attributes has been shown to depend on several contextual factors, such as the perceptual or categorical similarity with the task-relevant stimuli (Kahneman & Henik, 1981; Miles, Yamaguchi, & Proctor, 2009) and responses (Durgin, 2000). More relevant to the current study, Spruyt et al. (2009) used an affective priming task in which participants were required to evaluate either the affective value (positive vs. negative) or a non-affective value (human vs. object) of the target words in 75% of the trials. They found the affective priming effect when participants evaluated the affective value, but not when they evaluated the non-affective value. The results were replicated with masked primes (Spruyt, De Houwer, Everaert, & Hermans, 2012), indicating that this conditional priming effect did not depend on whether participants were aware of the primes. These studies suggest that automatic processing of irrelevant stimuli is conditional on the context or the requirement of the task (also see Yamaguchi & Proctor, 2011, 2012). Such conditional automaticity may explain the lack of

task-irrelevant IAT effects in the color task of Experiments 2 and 3. That is, task-irrelevant gender and racial attributes may not produce an IAT effect in the color task because participants are not prone to processing irrelevant social attributes of individuals when the task-relevant attribute is not socially relevant. Alternatively, processing of task-irrelevant attributes may depend on perceptual salience. Even if task-irrelevant attributes are processed, their influences may be a function of their relative salience, such that the task-irrelevant attributes exert stronger influences on responding when they are perceptually more salient. The present experiment aims to disentangle these possibilities.

The experiment was essentially the same as Experiment 2, in which participants performed the color task and the race task. In the color task of the present experiment, however, the color frame was removed from the display; instead, the eyes of individuals in the images were colored in green or red, and participants categorized the photographs based on the eye colors. The race task used the same photographs as those in the color task, but the eye colors were irrelevant. Our assumption was that the eye color will be perceived as a personal characteristic of an individual, in a similar fashion to their race and gender, because eye color is often considered to be an important quality that determines one's perceived attractiveness (Beattie & Shovelton, 2002; Grudl, Knoll, Eisenmann-Klein, & Prantl, 2012), group membership (Stewart, 2003), or even mate selection (Frost, 2006). If so, the conditional automaticity hypothesis would predict that the task-irrelevant social attributes should also be processed and the task-irrelevant race and gender IAT effects should be reinstated in the color task. In the race task, the task-irrelevant color and gender IAT effects should also be obtained, given that all of these attributes are socially relevant.

We also note that this assumption that eye color is perceived as a social attribute could be wrong. Instead, eye color may still be processed as a perceptual attribute as in the preceding experiments, and perceptual salience of attributes may determine the IAT effect, as the relative salience hypothesis proposes. The colors occupied small areas of the images as compared to the color frames in the previous two experiments, so the perceptual salience of color was made lower in the present experiment. In the race task, the relative salience hypothesis provides an unambiguous prediction that the influence of color in the race task should be reduced, unlike in Experiment 2, because the relative salience of color was made lower. Thus, the task-irrelevant color IAT effect may be eliminated in the race task. In the color task, on the other hand, one might well reason that the salience of task-irrelevant social attributes would increase relative to that of colors, so the IAT effects should be obtained based on these social attributes. Yet, one can also argue that participants would be forced to focus their attention on the small area of the photograph, which reduces attention to task-irrelevant social attributes and their influences. Therefore, the predictions of the relative salience hypothesis for the color task is ambiguous, and we remain neutral in this respect.

Consequently, if the conditional automaticity hypothesis is correct, the task-irrelevant IAT effects should now occur based on social attributes (race and gender) in the color task as well as eye colors in the race task. If the relative salience hypothesis is correct, the task-irrelevant color IAT effect should now disappear in the race task, regardless of whether task-irrelevant social attributes yield the IAT effect in the color task.

Method

Participants

Twenty four White females (mean age = 21.79, *SD* = 6.12, range = 18-44) were newly recruited from the same subject pool, with the same selection criteria as before.

Apparatus, Stimuli, and Procedure

The present experiment was essentially the same as Experiment 2, except for the way colors were presented. Unlike Experiment 2 in which the color of the picture frame was varied across trials, the present experiment removed the picture frame (as in Experiment 1) and varied the colors of the eyes on the faces. The face stimuli were the same as those used in Experiment 2, but the pupil was colored in red or green (the stimuli can be found in the OSF project page as noted in the author notes). We decided to use red and green to match with the colors used in Experiments 2 and 3, although they were not common colors in human eyes. In the color task, participants selected a response based on the pupil color, but they ignored the color in the race task. The procedure followed that of Experiment 2 in other respects.

Results

The data were filtered (0.93% of all trials) and analyzed in the same manner as in the preceding experiments. The overall error rate was 5.86%. Mean RTs were submitted to 2 (Race Compatibility) x 2 (Gender Compatibility) x 2 (Color Compatibility) ANOVAs separately for the race and color tasks. Color compatibility was coded as in Experiments 2 and 3. RT and PE are shown in Figure 4 and in Table 8. The results of ANOVA are summarized in Table 9.

Race task

For RT, there was a main effect of Race Compatibility, which showed an 80-ms advantage for the White-good/Black-bad mapping (*M* = 659 ms) than for the White-bad/Black-good mapping (*M* = 739 ms). There was also an interaction between Race Compatibility and Gender Compatibility. There was a gender IAT effect of 27 ms (*Ms* = 646 ms vs. 673 ms for

gender compatible and incompatible trials, respectively) when the race was compatible; there was a reversed gender IAT effect of –38 ms (*M*s = 758 ms vs. 720 ms for gender compatible and incompatible trials, respectively) when the race was incompatible. These outcomes were consistent with Experiment 1. But, unlike Experiment 1, the present experiment produced no influence of Color Compatibility. The results are consistent with the relative salience hypothesis but are inconsistent with the conditional automaticity hypothesis. PE did not show any significant effect.

Color task

For RT, the only significant effect was a main effect of Color Compatibility. Responses were faster for the green-good/red-bad mapping (*M* = 621 ms) than for the green-bad/red-good mapping (*M* = 720 ms). There was no influences of the gender or race when they were irrelevant to the task. The results are also inconsistent with the conditional automaticity hypothesis, whereas they neither support nor contradict the relative salience hypothesis.

For PE, Color Compatibility also showed a marginal effect; responses were more accurate for the green-good/red-bad mapping (*M* = 3.55%) than for the green-bad/red-good mapping (*M* = 5.73%). No other effects reached significance.

Discussion

The present results provided little support for the conditional automaticity hypothesis. Instead, the outcomes of the race task were consistent with the relative salience hypothesis. As the colors occupied smaller areas of the images, the colors were no longer salient enough to exert an influence when they were task-irrelevant in the race task. These results point to the role of attention in producing the task-irrelevant IAT effects, which also raises a question about the implicitness of the task-irrelevant attributes that produce the IAT effect. Note that the present

results corroborate the conclusion in Experiment 3 that the racial IAT effect is not due to a perceptual quality of individuals (e.g., skin color), given that the race IAT only produced the gender effect but not the color effect, whereas the color IAT only produced the color effect but not the race effect. These findings have important implications as to the origin of racial IAT effects.

General Discussion

In the IAT, participants are asked to categorize material based upon particular attributes of interest. For example, to measure implicit attitudes toward race, participants are required to categorize individuals according to their race. It is unclear as to how such task requirements may impact upon participants' perception of the material in question, and this raises an important question as to whether implicit attitudes toward attributes of an individual in the target stimulus influence participants' judgement when these attributes are actually *irrelevant* to performing the task. To address this issue, we used the m-IAT that involved multiple social or non-social attributes, some of which were relevant and some of which were irrelevant to the task in hand. By using this paradigm, the present study asked the following questions: firstly whether social biases that are measured in the standard IAT are still observed when the attribute in question (e.g. race or gender) is irrelevant to the performance of the task, and secondly how one's intention to categorize any attribute influences the categorization of other attributes. In the following sections, we first summarize the findings from the four experiments and consider their theoretical implications. We then consider the societal significance of our results.

Theoretical Implications and Remaining Empirical Issues

Experiment 1 showed that both race and gender biases do influence people's categorization performance even when these attributes are irrelevant to the task in hand.

Nevertheless, the effects of task-irrelevant attributes were smaller than the effect of the task-relevant attribute, and the direction of the task-irrelevant IAT effect depended on whether the task requires participants to counter against the bias based on the task-relevant attribute. Thus, the task-irrelevant IAT effect was reversed when categories were assigned to responses incompatible with participants' biases. These results corroborate the classic finding of Hedge and Marsh (1975) in the Simon task, where the Simon effect is reversed when participants are required to select response colors that are incompatible with stimulus colors. Experiment 2 showed that more than one task-irrelevant attribute could influence categorization performance at the same time, and that countering a bias based on a task-irrelevant attribute has no influence on the task-irrelevant IAT effect based on another task-irrelevant attribute. The results excluded the possibility that the IAT effect reversed when two implicit associations were inconsistent. Instead, overcoming one implicit bias reverses the effects of implicit biases in other domains, but the reversal requires an intention to overcome bias, as overcoming a bias on a task-irrelevant attribute does not alter the influence of other task-irrelevant attributes. Therefore, the results highlight the importance of the intention to overcome social biases in modulating the influences of implicit attitudes.

Experiments 3 and 4 followed up these intriguing findings of Experiment 2 and showed that performing a categorization based on color did not produce any influence of task-irrelevant attributes of race and gender, whereas performing the categorization of race did produce an influence of the task-irrelevant color attribute. These asymmetric results are suggestive of how explicit categorization determines the influences of task-irrelevant attributes. We considered three possible hypotheses which may help us understand the processes underpinning this: the relative speed hypothesis, the conditional automaticity hypothesis, and the relative salience

hypothesis. Experiment 3 tested the relative speed hypothesis by varying the interval between the onset of target colors and facial images. Although the relative speed hypothesis predicted that the task-irrelevant IAT effects based on race and gender would be obtained when the facial images were presented before the target color, the results did not support this. Experiment 4 tested the conditional automaticity hypothesis and the relative salience hypothesis. The conditional automaticity hypothesis predicted that the task-irrelevant IAT effects based on gender and race would be obtained if the colors were presented as part of the individuals' personal characteristics. The relative salience hypothesis predicted that the task-irrelevant IAT effects depend on the availability of attention to perceptually salient attributes, and that task-irrelevant IAT effect based on color should disappear when the colors were made less salient. Our results were consistent with the relative salience hypothesis.

The results supporting the relative speed hypothesis also corroborate the reversal of the Simon effect found in Hedge and Marsh's (1975) task, in which the irrelevant stimulus attribute (spatial position) is much different from the target attribute (color). Studies have shown that encoding of spatial positions are obligatory (e.g., Logan, 1998), even though the target attribute is non-spatial. Hence, the irrelevant stimulus attribute is salient in Hedge and Marsh's task, as the colors were in the present experiment; other social attributes of individuals might not have been sufficiently salient as compared to colors. Nevertheless, we also note the possibility that our assumption of eye colors being processed as a social cue may be incorrect in the first place (especially, given that we used red and green as eye colors, instead of more natural eye colors, in order to match the color of the stimuli used in the preceding experiments). We must acknowledge that our results do not refute the conditional automaticity hypothesis entirely and that a stronger test of the conditional automaticity hypothesis will be required in future

investigations. Indeed, we would suggest that the relative salience hypothesis and the conditional automaticity hypotheses are not, in fact, mutually exclusive, and they might both be correct at the same time.

In summary, the present set of experiments imply that the influence of task-irrelevant attributes is contingent on categorization of the task-relevant attribute and that biases measured in the IAT may not be implicit to participants who are performing the task. Also, a corollary to these findings is that the racial IAT effect is not a perceptually-based effect (such as the perceptual quality of individuals' skin colors) because the race effect and color effect do not coincide in the current variants of the IAT; if both effects depended on perceptual properties, they should co-occur.

Societal Significance and Implications

The findings of the present studies shed new light on the cognitive processes underlying the manifestation of social biases, and they may have far-reaching implications regarding the very nature of social biases and their possible effects on social judgements, as well as their relevance for social equality and social justice. We consider five possible implications.

Firstly, our studies have shown that social biases can affect people's judgment of others not only on the basis of the exact characteristics that are under consideration at any given time, but also on the basis of characteristics that are *irrelevant* to the matter at hand. Thus, these experiments confirm that social biases can affect judgments when they are not relevant to the judgment being made. This has not been previously been demonstrated in research using the IAT, although our conclusion is, in fact, corroborated by the findings in other tasks, such as the pronunciation task, in which participants are not required to evaluate any stimuli but produce in-

group advantage in reading affective words (e.g., Bargh et al., 1996). We should also note that the influence of irrelevant attributes depends on whether these attributes are sufficiently salient.

Secondly, we found, in line with all previous IAT studies, that it is possible to counter one's own social bias by intention (otherwise, correct responses would rarely occur in the incompatible blocks of the IAT). Our results also demonstrated that such intention to overcome a bias can influence the effects of other biases that one might have about other social characteristics of the target individuals. For instance, we found that the intention to counter a racial bias can eliminate or even reverse the IAT effect based on a gender bias. It does require a little caution to interpret this reversal as necessarily reflecting the actual reversal of the underlying bias itself, because our results only showed a reversal of the 'effect' of a social bias and there could be reasons other than a reversal of the bias itself that result in a reversed effect. An alternative explanation would be that the reversal reflected the congruence between two biases on a given trial. Responses might be faster when two biases are both compatible or incompatible at the same time than when one is compatible and the other is incompatible. Nevertheless, this account does not seem to fit the finding in Experiment 2 that two task-irrelevant biases did not reverse the IAT effects, indicating that mere congruence between social biases does not explain the results obtained. Further studies are clearly required to elucidate the mechanisms of this reversal.

Thirdly, the lack of an influence of a task-irrelevant bias on another task-irrelevant bias implies that responding against a task-irrelevant bias does not modulate the influence of another social bias. This finding is important. For example, when an intervention program is designed to eliminate one type of social bias, the elimination, or reversal, of this bias may not transfer automatically to eradicate another type of social bias, unless the trainee continues to suppress

their own bias intentionally. As previous studies of transfer of learning has shown (e.g., Proctor, Yamaguchi, Zang, & Vu, 2009; Yamaguchi & Proctor, 2009; Yamaguchi et al., 2015), the effectiveness of bias modification training would be limited to the very attribute that trainees are being trained for. The extent to which the effect of training transfers to untrained social biases would be an interesting topic to explore in future investigations.

Fourthly, our study also suggests that paradigms that are meant to measure implicit biases may only be applicable to the dimensions that are related closely to the type of judgments required in the paradigms used. That is, the IAT may be used to measure an irrelevant racial bias when the task requires subjects to categorize individuals based on another *social* attribute of individuals, but it may not work when it requires subjects to categorize individuals based on a *non-social* attribute, such as the dresses that the individuals are wearing (e.g., in terms of style or appropriateness; Spruyt et al., 2009). There have been a number of significant and vigorous debates as to the validity of the race IAT in predicting discrimination behaviors (e.g., Carlsson & Ageström, 2016; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; McConnell & Leibold, 2001; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), and the results do look somewhat mixed. The present series of studies provide new evidence about some important aspects of when and how task-irrelevant IAT results emerge, which not only helps us understand discrepancies in the literature but also provides a new theoretical platform for future research in this area. A corollary to the present finding is that race is processed differently from color, implying that the racial effect is not based on a perceptual quality of individuals; that is, it is not the skin color itself that is producing the racial IAT effect, but it is what the skin color represents – it is the race of a given individual that is the origin of the racial IAT effect. Hence, the dissociation

between the racial and color IAT effects in the present study provides an important contribution to the debate as to the origins of the IAT effect.

Finally, we should note that, as shown in previous studies (e.g., Aidman & Carroll, 2003; Meissner & Rothermund, 2013; Mitchell et al., 2003; Nosek & Banaji, 2003), the gender IAT effect appears more unambiguous for female participants than for male participants (Experiment 1). Given that the main purpose of the present research required more than two attributes to produce the IAT effect, we decided to focus on female participants in the subsequent experiments. This may, however, limit the generalizability of the present set of results. We also note here that all four experiments took place in a university campus as do most psychological studies. Although we do not have strong reasons to doubt the generalizability of the findings beyond our participant population, this remains an empirical issue of potentially enormous societal importance that awaits scrutiny in different locations and with different populations.

Concluding Remarks

We have demonstrated that implicit biases (related to both race and gender) do still emerge when the attribute is irrelevant to the task in hand, but the biases which emerge are smaller in magnitude. Our results also indicated that the manifestations of such implicit biases depend on explicit categorization of the task-relevant attribute categories that participants must be aware of in order to complete the task accurately. We have attempted to uncover the processes operating in the IAT and forge connections with theoretical ideas from mainstream cognitive psychology, including research on the Simon effect and the Stroop effect, to cast new light on these processes. We suggest that this set of studies will necessitate a rethink of how we define implicit processes in the IAT and should lead to a more critical and focussed perspective on race and gender bias research going forward.

Context

Given the current climate in globalized communities in North America, Europe, and elsewhere, it is crucial that individuals are treated in a fair manner, regardless of their origin, race, gender, and other salient characteristics that are often irrelevant to selection processes for occupations or education. However, it has been very difficult to determine how potential biases based on irrelevant personal characteristics could be measured. Although the implicit association test (IAT) has been very popular, there has been considerable debate as to the origin of the effects measured in the test. The present article provides an experimental study that investigated cognitive bases of the IAT effect and revealed that the explicit categorization required in the task plays a significant role in determining the effect size. The results also indicate that the occurrence of the IAT effect depends on the kinds of category used in the task.

References

Aidman, E. V., & Carroll, S. M. (2003). Implicit individual differences: Relationship between implicit self-esteem identity, and gender attitude. *European Journal of Personality*, *17*, 19-37.

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A handbook of social psychology* (*pp*. 798-844). Worcester, MA: Clark University Press.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the implicit association test?" *Psychological Inquiry*, *15*, 257-278.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face images. *Journal of Experimental Psychology: General*, *142*, 1323-1334.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavioral Research, 46*, 668-688.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3-51). New York: Guilford.

Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditional automatic attitude activation with a pronunciation task. *Journal of Experimental Social Psychology*, *32*, 104-128.

Baroni, G., Yamaguchi, M., Chen, J., & Proctor, R. W. (2013). Mechanisms underlying transfer of task-defined rules across feature dimensions. *Experimental Psychology*, *60*, 410-424.

Beattie, G. (2010) *Why Aren't We Saving the Planet? A Psychologist's Perspective*. London: Routledge.

Beattie, G. (2013) *Our Racist Heart? An Exploration of Unconscious Prejudice in Everyday life.* London: Routledge.

Beattie, G., Cohen, D. L., & McGuire, L. (2013). An exploration of possible unconscious ethnic biases in higher education: The role of implicit attitudes on selection for university posts. *Semiotica, 197,* 217-247.

Beattie, G. & Shovelton, H. (2002) Blue-eyed boys? A winning smile? An experimental investigation of some core facial stimuli that may affect interpersonal perception. *Semiotica, 139,* 1-21.

Berger, J. (2018). Implicit attitudes and awareness. *Synthese, in press*.

Blanton, H., Jaccard, J., & Burrows, C. N. (2015). Implications of the implicit association test D-transformation for psychological assessment. *Assessment, 22,* 429-440.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Docoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42,* 192-212.

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology, 42,* 163-176.

Carlsson, R., & Ageström, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian Journal of Psychology, 57,* 278-287.

Chang, B. P., & Mitchell, C. J. (2011). Discriminating between the effects of valence and salience in the implicit association test. *Quarterly Journal of Experimental Psychology, 64,* 2251-2275.

Conray, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469-487.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*, 443-451.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50*, 77-85.

De Houwer, J., & Eelen, P. (1998). An affective variant of the Simon paradigm. *Cognition and Emotion, 12*, 45-62.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*, 347-368.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5-18.

Downs, A. C., & Lyons, P. M. (1991). Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin, 17*, 541-547.

Durgin, F. H. (2000). The reverse Stroop effect. *Psychonomic Bulletin & Review, 7*, 121-125.

Fazio, R. H. (1989). On the power and functionality of attitudes: The role of attitude accessibility. In A. R. Pratkanis, S. J. Breckler & A. G. Greenwald (Eds), *Attitude structure and function*. Hillsdale, NJ: Erlbaum.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the implicit association test (IAT). *European Review of Social Psychology, 17*, 74-147.

Fazio, R.H., Sanbonmatsu, D.M., Powell, M.C., & Kardes, F.R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238.

Frost, P. (2006). European hair and eye colour: A case of frequency-dependent sexual selection? *Evolution & Human Behavior, 27,* 85–103

Gawronski, B., & Bodenhausen, G. V. (2011). Accessibility effects on implicit social cognition: The role of knowledge activation and retrieval experiences. *Journal of Personality and Social Psychology, 89,* 672-685.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 875-8794.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40,* 357-365.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17-41.

Gründl, M., Knoll, S., Eisenmann-Klein, M., & Prantl, L. (2012). The blue-eyes stereotype: Do eye colour, pupil diameter, and scleral colour affect attractiveness? *Aesthetic Plastic Surgery, 36,* 234-240.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*, 1369-1392.

Hedge, A., & Marsh, N. W. A. (1975). The effect of irrelevant spatial correspondences on two-choice response-time. *Acta Psychologica*, *39*, 427-439.

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J.R. Pomerantz (Eds.), Perceptual organization (pp. 181-211). Hillsdale, NJ: Erlbaum.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D R. Davies (EDs.), *Varieties of attention* (pp. 29-61). Orlando, FL: Academic Press.

Klauer, K. C. (1997). Affective priming. *European Review of Social Psychology*, *8*, 67-103.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility – A model and taxonomy. *Psychological Review*, *97*, 253-270.

Koole, S. L., Dijksterhuis, A., & van Knippenberg, A. (2001). What's in a name: Implicit self-esteem and the automatic self. *Journal of Personality and Social Psychology*, *80*, 669-685.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology*, *82*, 774-778.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492-527.

Logan, G. D. (1998). What is learned during automatization? II: Obligatory encoding of location information. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1720-1736.

Lu, C.-H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review, 2*, 174–207.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Review*, *109*, 163-203.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*, 435-442.

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the implicit association test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, *104*, 45-69.

Melnikoff, D., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, *22*, 280-293.

Miles, J. D., Yamaguchi, M., & Proctor, R. W. (2009). Dilution of compatibility effects in Simon-type tasks depends on categorical similarity between distractors and diluters. *Attention, Perception, & Psychophysics*, *71*, 1598-1606.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455-469.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*, 297-326.

Notebaert, W., Gevers, W., Verguts, T., & Fias, W. (2006). Shared spatial representations for numbers and space: The reversal of the SNARC and the Simon effects. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 1197-1207.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition¸ 19*, 625-666.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences*, *15*, 152-159.

Nosek, B. A., & Sriram, N. (2007). Faulty assumptions: A comment on Blanton, Jaccard, Gonzales, & Christie (2006). *Journal of Experimental Social Psychology*, *43*, 393-398.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*, 636-639.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171-192.

Payne, B.K., Cheng, C. M., Govorun, O., Stewart, B. D. (2005) An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277–293.

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*, 187-255.

Porter, S., ten Brinke, L., & Gustaw, C. (1991). Dangerous decisions: The impact of first impression of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime, & Law*, *16*, 477-491.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 153-175). Hillsdale, NJ: Erlbaum.

Proctor, R. W., Yamaguchi, M., Zhang, Y., & Vu, K. P.-L. (2009). Influence of visual stimulus mode on transfer of acquired spatial associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 434-445.

Sarnoff, I. (1960). Psychoanalytic theory and social attitudes. *Public Opinion Quarterly*, *24*, 251-279.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*, 1-66.

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*, 752-775.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, *115*, 314-335.

Steffens, M., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Experimental Psychology*, *48*, 123-134.

Stewart, T. (2003). Do the "Eyes" have it? A program evaluation of Jane Elliott's "Blue-Eyes/Brown-Eyes" diversity training exercise. *Journal of Applied Social Psychology*, *33*, 1898-1921.

Spruyt, A., De Houwer, J., Everaert, T., & Hermans, D. (2012). Unconscious semantic activation depends on feature-specific attention allocation. *Cognition*, *122*, 91-95.

Spruyt, A., De Houwer, J., & Hermans, D. (2009). Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory and Language*, *61*, 37-54.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25-29.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529-554.

Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naïve theories in corrections for perceived bias. *Journal of Personality and Social Psychology*, *68*, 36-51.

Yamaguchi, M., Chen, J., & Proctor, R. W. (2015). Transfer of learning in choice reactions: The roles of stimulus type, response mode, and set-level compatibility. *Memory & Cognition*, *43*, 825-836.

Yamaguchi, M., & Proctor, R. W. (2009). Transfer of learning in choice-reactions: Contributions of specific and general components of manual responses. *Acta Psychologica*, *130*, 1-10.

Yamaguchi, M., & Proctor, R. W. (2011). The Simon task with multi-component responses: Two loci of response-effect compatibility. *Psychological Research*, *75*, 214-226.

Yamaguchi, M., & Proctor, R. W. (2012). Multidimensional vector model of stimulus-response compatibility. *Psychological Review*, *119*, 272-303.

Table 1. Mean ratings of five attributes of the face stimuli used in Experiment 1 (values in the parentheses are standard deviations).

	Black Female		Black Male		White Female		White Male	
Attractiveness	3.00	(0.53)	3.00	(0.41)	3.01	(0.46)	3.00	(0.43)
Typicality	3.11	(0.33)	3.04	(0.31)	3.15	(0.31)	3.10	(0.30)
Strength of Emotion	3.54	(0.58)	3.41	(0.83)	3.39	(0.44)	3.41	(0.76)
Friendliness	3.92	(0.46)	3.77	(0.66)	3.80	(0.39)	3.85	(0.46)
Memorability	3.11	(0.37)	3.17	(0.41)	3.00	(0.34)	3.16	(0.39)

Table 2. Percentage Errors in Experiment 1.

Participants	Task	Gender Compatibility	Race Compatibility			
			Compatible		Incompatible	
Female	Race	Compatible	4.53	(.78)	4.28	(.86)
		Incompatible	5.05	(.79)	4.61	(.92)
	Gender	Compatible	5.55	(.79)	6.80	(1.08)
		Incompatible	10.01	(1.35)	7.38	(1.14)
Male	Race	Compatible	1.99	(.45)	3.87	(.82)
		Incompatible	3.66	(.62)	3.34	(.79)
	Gender	Compatible	4.17	(.76)	6.63	(1.00)
		Incompatible	3.35	(.60)	4.65	(.79)

Table 3. Results of the Analyses of Variance on Response Time (RT) and Percentage Errors (PE) in Experiment 1.

Factor	*df*	*MSE*	*F*	*p*	η_p^2
		Females: RT			
Task	1,31	7969.23	2.30	.139	.069
Gender Compatibility (GC)	**1,31**	**3549.14**	**42.23**	**< .001**	**.577**
Race Compatibility (RC)	**1,31**	**5434.76**	**11.41**	**.002**	**.269**
Task x GC	**1,31**	**2749.97**	**17.21**	**< .001**	**.357**
Task x RC	**1,31**	**6038.42**	**6.28**	**.018**	**.168**
GC x RC	**1,31**	**2315.94**	**28.21**	**< .001**	**.476**
Task x GC x RC	1,31	1702.73	< 1	.742	.004
		Females: PE			
Task	**1,31**	**22.18**	**22.91**	**< .001**	**.425**
GC	**1,31**	**25.51**	**5.45**	**.026**	**.149**
RC	1,31	26.75	< 1	.430	.020
Task x GC	1,31	18.05	3.92	.057	.112
Task x RC	1,31	19.46	< 1	.752	.003
GC x RC	1,31	23.87	2.76	.107	.082
Task x GC x RC	1,31	17.25	3.17	.085	.093
		Males: RT			
Task	1,31	12849.57	< 1	.456	.018
GC	1,31	5445.73	4.05	.053	.116
RC	**1,31**	**5174.69**	**40.52**	**< .001**	**.567**
Task x GC	1,31	6696.03	4.00	.054	.114
Task x RC	**1,31**	**5457.95**	**21.96**	**< .001**	**.415**
GC x RC	1,31	2206.35	< 1	.461	.018
Task x GC x RC	1,31	2412.78	2.52	.123	.075
		Males: PE			
Task	**1,31**	**13.50**	**10.44**	**.003**	**.252**
GC	1,31	19.84	< 1	.460	.018
RC	**1,31**	**15.99**	**7.09**	**.012**	**.186**
Task x GC	**1,31**	**10.53**	**5.86**	**.022**	**.159**
Task x RC	1,31	13.76	1.41	.245	.043
GC x RC	**1,31**	**8.51**	**5.31**	**.028**	**.146**
Task x GC x RC	1,31	7.75	< 1	.459	.018

Note: Bold indicates significant effects at alpha = .05.

Table 4. Percentage Errors in Experiment 2.

Task	Color Compatibility	Gender Compatibility	Race Compatibility			
			Compatible		Incompatible	
Race	Compatible	Compatible	5.00	(1.40)	7.03	(1.21)
		Incompatible	2.75	(.89)	8.25	(1.93)
	Incompatible	Compatible	8.25	(1.60)	8.53	(2.01)
		Incompatible	7.33	(1.45)	6.57	(1.79)
Color	Compatible	Compatible	4.25	(.86)	5.78	(.96)
		Incompatible	2.88	(.67)	5.88	(1.27)
	Incompatible	Compatible	8.78	(1.29)	7.92	(1.39)
		Incompatible	6.56	(1.23)	6.42	(1.14)

Table 5. Results of the Analyses of Variance on Response Time (RT) and Percentage Errors (PE) in Experiment 2.

Factor	*df*	*MSE*	*F*	*p*	η_p^2
		Race Task: RT			
Color Compatibility (CC)	1, 24	1693.84	< 1	.684	.007
Gender Compatibility (GC)	1, 24	2486.43	1.07	.311	.043
Race Compatibility (RC)	**1, 24**	**15918.67**	**11.49**	**.002**	**.324**
CC x GC	1, 24	3571.58	< 1	.425	.027
CC x RC	**1, 24**	**5587.92**	**6.46**	**.018**	**.212**
GC x RC	**1, 24**	**3287.34**	**15.67**	**.001**	**.395**
CCx GC x RC	1, 24	3996.00	< 1	.802	.003
		Race Task: PE			
CC	1, 24	56.18	3.26	.084	.119
GC	1, 24	34.13	1.41	.248	.055
RC	1, 24	70.89	2.19	.152	.084
CC x GC	1, 24	42.08	< 1	.619	.010
CC x RC	1, 24	52.18	3.85	.061	.138
GC x RC	1, 24	49.58	< 1	.550	.015
CCx GC x RC	1, 24	25.11	2.54	.124	.096
		Color Task: RT			
CC	**1, 24**	**16896.48**	**28.20**	**< .001**	**.540**
GC	1, 24	3420.56	1.91	.180	.074
RC	1, 24	2552.82	< 1	.393	.031
CC x GC	1, 24	3855.17	< 1	.670	.008
CC x RC	1, 24	5359.09	1.62	.215	.063
GC x RC	1, 24	2699.71	< 1	.698	.006
CCx GC x RC	1, 24	2734.75	< 1	.499	.019
		Color Task: PE			
CC	**1, 24**	**29.49**	**12.57**	**.002**	**.344**
GC	**1, 24**	**11.68**	**6.68**	**.016**	**.218**
RC	1, 24	16.78	2.31	.142	.088
CC x GC	1, 24	32.01	< 1	.452	.024
CC x RC	1, 24	28.55	3.33	.081	.122
GC x RC	1, 24	20.07	< 1	.398	.030
CCx GC x RC	1, 24	19.96	< 1	.768	.004

Note: Bold indicates significant effects at alpha = .05.

Table 6. Percentage Errors in Experiment 3 (COD = Color Onset Delay in millisecond).

COD	Color Compatibility	Gender Compatibility	Race Compatibility			
			Compatible		Incompatible	
100	Compatible	Compatible	2.43	(.81)	2.56	(.95)
		Incompatible	3.13	(.93)	3.26	(1.06)
	Incompatible	Compatible	4.38	(1.05)	4.08	(.94)
		Incompatible	4.54	(1.02)	3.14	(.84)
250	Compatible	Compatible	1.52	(.72)	2.21	(.77)
		Incompatible	1.82	(.52)	2.47	(.81)
	Incompatible	Compatible	4.20	(.91)	4.36	(.96)
		Incompatible	2.30	(.57)	2.97	(.89)
500	Compatible	Compatible	1.62	(.64)	1.79	(.72)
		Incompatible	2.35	(.81)	2.45	(.73)
	Incompatible	Compatible	4.19	(1.00)	3.01	(1.06)
		Incompatible	4.37	(1.00)	3.32	(.87)

Table 7. Results of the Analyses of Variance on Response Time (RT) and Percentage Errors (PE) in Experiment 3.

Factor	*df*	*MSE*	*F*	*p*	η_p^2
			RT		
Color Onset Delay (COD)	**2,46**	**2334.09**	**78.57**	**< .001**	**.774**
Color Compatibility (CC)	**1,23**	**46369.50**	**27.93**	**< .001**	**.548**
Gender Compatibility (GC)	1,23	2705.14	< 1	.797	.003
Race Compatibility (RC)	1,23	2505.76	< 1	.425	.028
COD x CC	2,46	1893.50	< 1	.392	.040
COD x GC	2,46	2020.94	< 1	.560	.025
CC x GC	1,23	1995.00	< 1	.915	.001
COD x CC x GC	2,46	1458.01	1.20	.311	.050
COD x RC	1,23	1589.01	< 1	.998	< .001
CC x RC	1,23	2015.40	< 1	.717	.006
COD x CC x RC	2,46	1633.71	1.65	.203	.067
GC x RC	1,23	2853.58	< 1	.565	.015
COD x GC x RC	2,46	1823.87	< 1	.407	.038
CC x GC x RC	1,23	1132.96	2.07	.164	.083
COD x CC x GC x RC	2,46	2042.15	1.23	.302	.051
			PE		
COD	2,46	11.93	2.23	.120	.088
CC	1,23	94.08	3.17	.088	.121
GC	1,23	7.06	< 1	.926	< .001
RC	1,23	11.65	< 1	.726	.005
COD x CC	2,46	15.25	< 1	.834	.008
COD x GC	2,46	12.38	1.37	.265	.056
CC x GC	1,23	14.72	3.22	.086	.123
COD x CC x GC	2,46	11.65	< 1	.574	.024
COD x RC	1,23	12.68	1.20	.310	.050
CC x RC	1,23	10.95	2.26	.146	.090
COD x CC x RC	2,46	16.19	< 1	.822	.009
GC x RC	1,23	14.75	< 1	.880	.001
COD x GC x RC	2,46	11.65	< 1	.844	.007
CC x GC x RC	1,23	9.71	< 1	.909	.001
COD x CC x GC x RC	2,46	14.17	< 1	.853	.007

Note: Bold indicates significant effects at alpha = .05.

Table 8. Percentage Errors in Experiment 4.

Task	Color Compatibility	Gender Compatibility	Race Compatibility			
			Compatible		Incompatible	
Race	Compatible	Compatible	7.17	(2.17)	6.88	(1.58)
		Incompatible	9.11	(2.84)	7.14	(1.48)
	Incompatible	Compatible	7.86	(3.18)	6.04	(1.33)
		Incompatible	8.37	(2.63)	4.26	(1.14)
Color	Compatible	Compatible	2.60	(.83)	3.98	(1.21)
		Incompatible	3.40	(1.13)	4.22	(1.11)
	Incompatible	Compatible	6.01	(1.48)	5.52	(1.22)
		Incompatible	5.82	(1.47)	5.58	(1.54)

Table 9. Results of the Analyses of Variance on Response Time (RT) and Percentage Errors (PE) in Experiment 4.

Factor	*df*	*MSE*	*F*	*p*	η_p^2
		Race Task: RT			
Color Compatibility (CC)	1,23	2536.41	< 1	.562	.015
Gender Compatibility (GC)	1,23	4336.89	< 1	.566	.015
Race Compatibility (RC)	**1,23**	**36696.87**	**8.35**	**.008**	**.266**
CC x GC	1,23	6020.51	< 1	.832	.002
CC x RC	1,23	4504.26	.11	.744	.005
GC x RC	**1,23**	**7474.73**	**6.65**	**.017**	**.224**
CCx GC x RC	1,23	3879.34	1.20	.284	.050
		Race Task: PE			
CC	1,23	51.49	< 1	.373	.035
GC	1,23	17.69	< 1	.707	.006
RC	1,23	310.38	< 1	.428	.028
CC x GC	1,23	15.03	2.42	.133	.095
CC x RC	1,23	46.48	< 1	.362	.036
GC x RC	1,23	28.56	1.66	.211	.067
CCx GC x RC	1,23	53.81	< 1	.888	.001
		Color Task:RT			
CC	**1,23**	**31590.90**	**14.75**	**.001**	**.391**
GC	1,23	6394.90	< 1	.935	< .001
RC	1,23	3157.07	1.39	.250	.057
CC x GC	1,23	4864.39	< 1	.332	.041
CC x RC	1,23	3039.99	< 1	.646	.009
GC x RC	1,23	6453.81	1.16	.293	.048
CCx GC x RC	1,23	5150.06	< 1	.612	.011
		Color Task:PE			
CC	1,23	55.28	4.13	.054	.152
GC	1,23	25.06	< 1	.754	.004
RC	1,23	39.97	< 1	.693	.007
CC x GC	1,23	24.84	< 1	.687	.007
CC x RC	1,23	44.58	< 1	.457	.024
GC x RC	1,23	27.57	< 1	.919	< .001
CCx GC x RC	1,23	36.14	< 1	.821	.002

Note: Bold indicates significant effects at alpha = .05.

Figure 1. Mean response times for female (A and B) and male (C and D) participants in the Gender and Race IAT tasks; error bars are standard errors of the means.

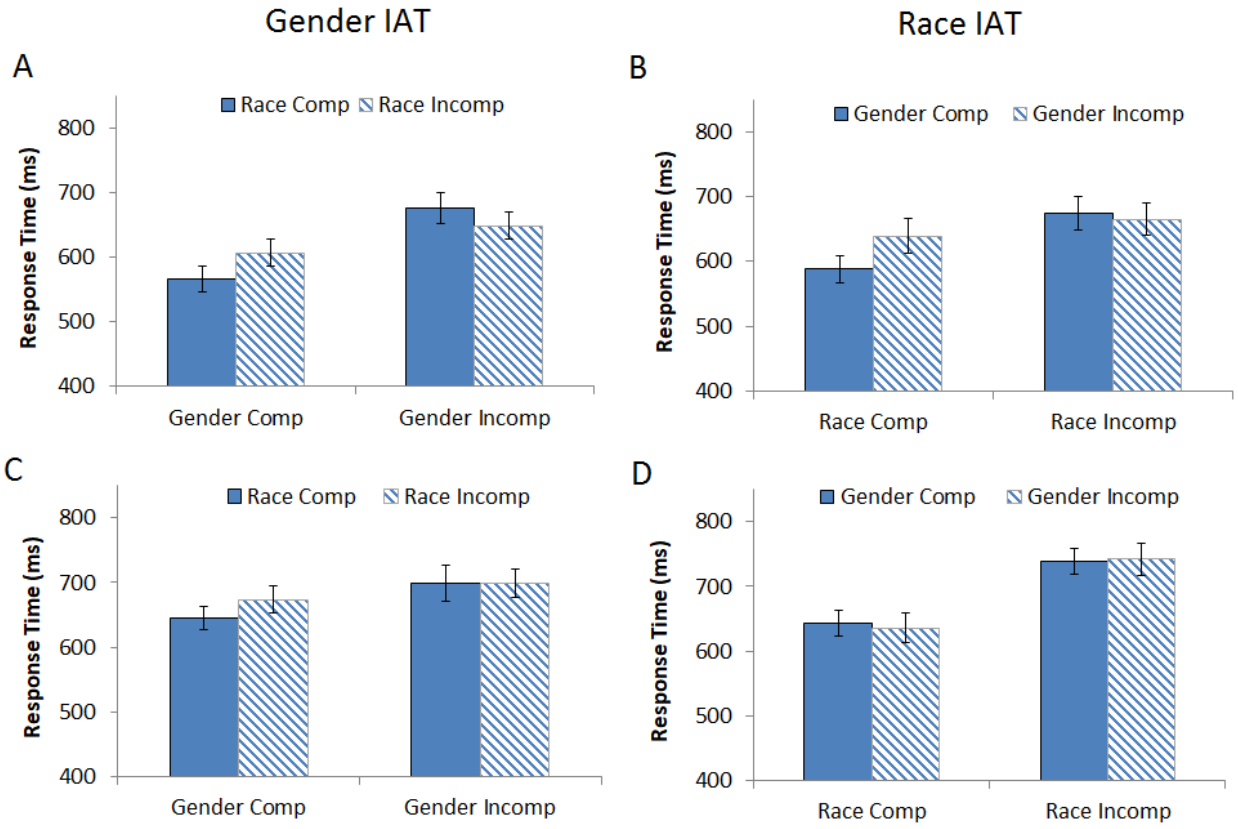

Figure 2. Mean response times in the Race IAT task (A) and the Color IAT task (B); error bars are standard errors of the means.

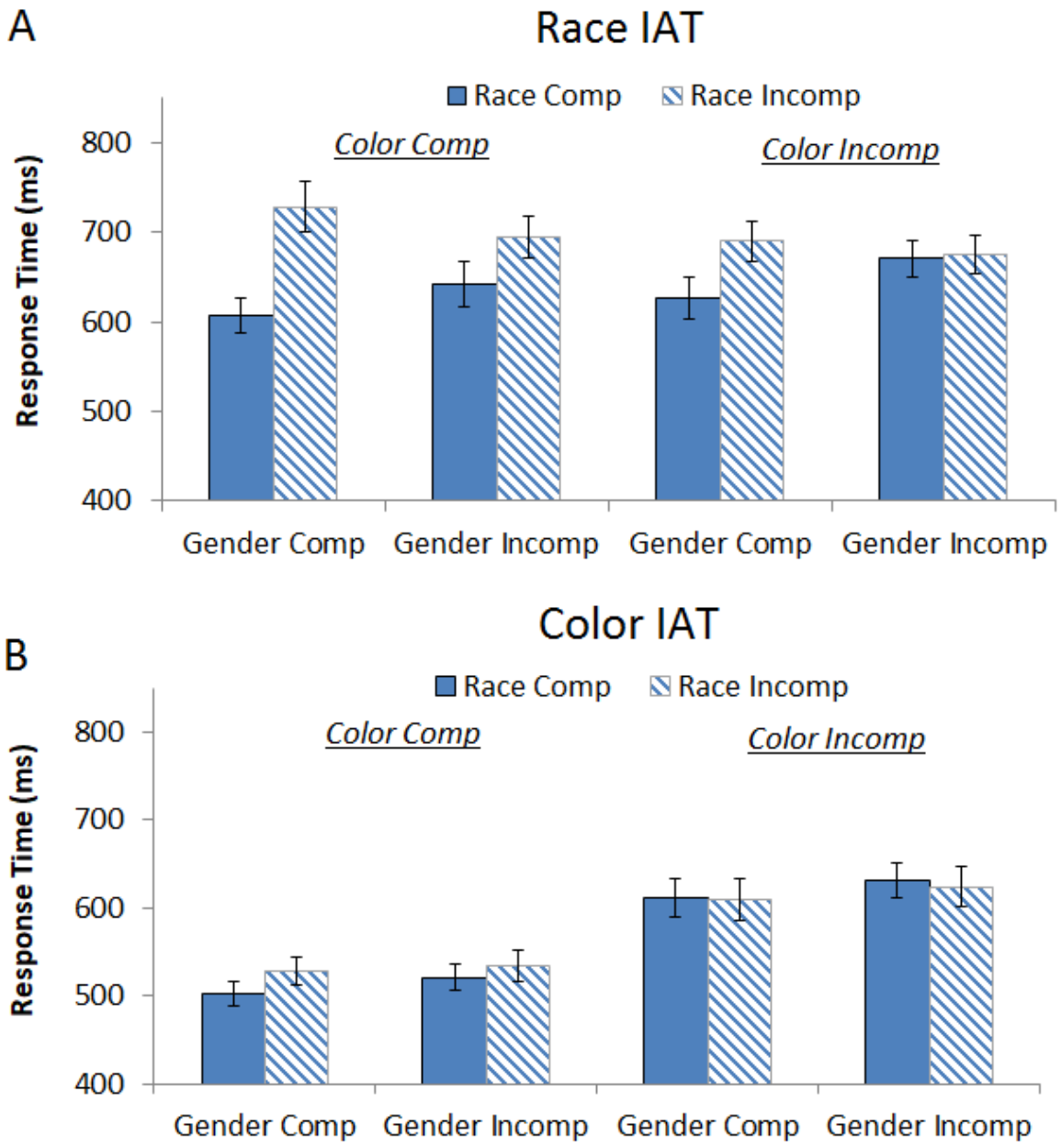

Figure 3. Mean response times in the Color IAT task as a function of SOA and Race Compatibility (A and B) or Gender Compatibility (C and D) for color compatible and incompatible trials, and as a function of SOA and Color Compatibility (E); error bars are standard errors of the means.

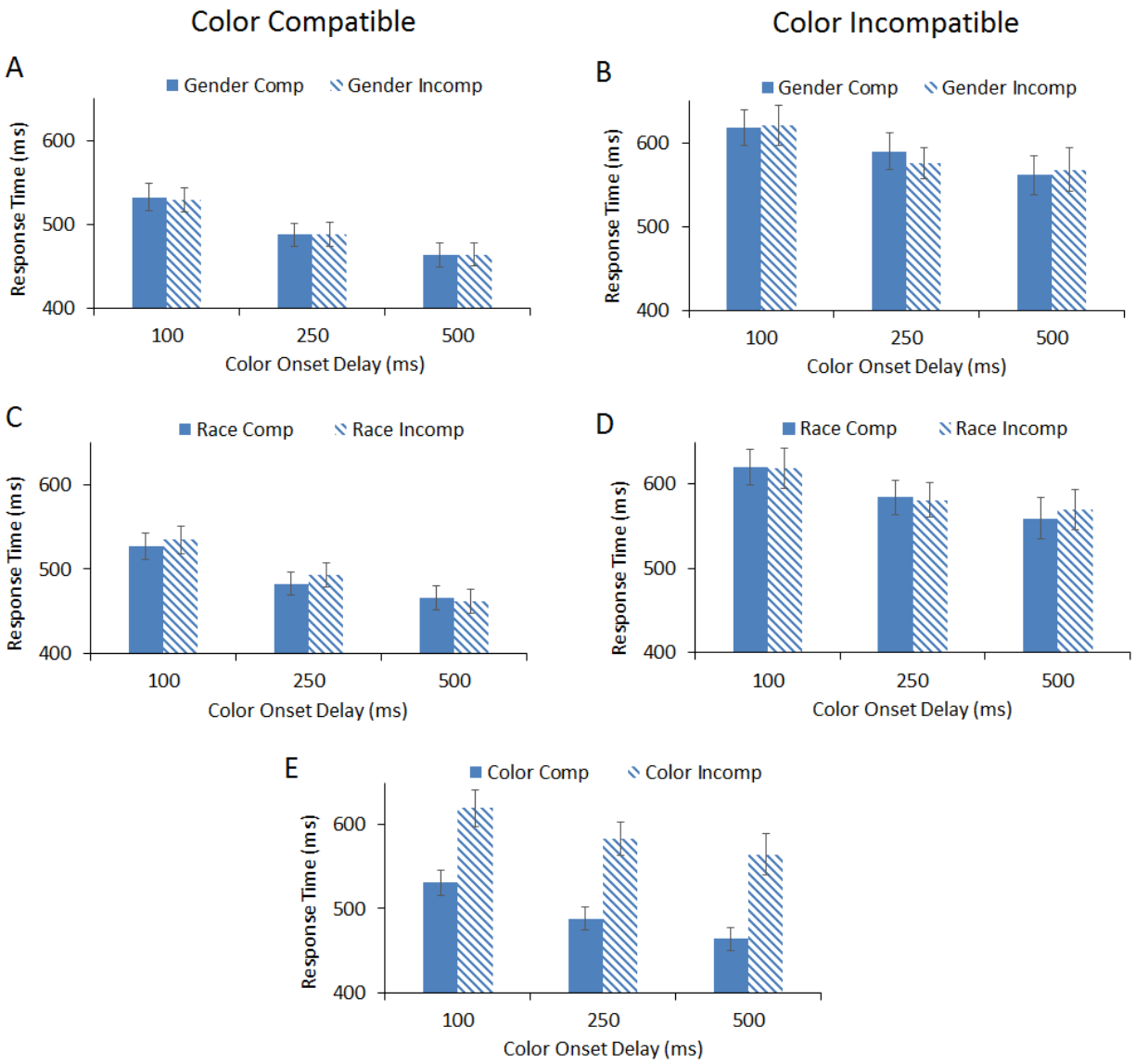

Figure 4. Mean response times in the Race IAT task (A) and the Color IAT task (B); error bars are standard errors of the means.