The Influence of Race on Weapon Identification: Cognitive Processes Underlying the Weapon Identification Task

---

submitted by

Ruben Laukenmann

from

Kleinansbach

Inaugural Dissertation

submitted in partial fulfillment of the requirements

for the degree Doctor of Social Sciences

in the DFG Research Training Group "Statistical Modeling in Psychology"

at the University of Mannheim

Thesis Defense:

15.12.2023

Supervisors:

Prof. Dr. Edgar Erdfelder

Prof. Dr. Thorsten Meiser

Dean of the School of Social Sciences:

Prof. Dr. Michael Diehl

Evaluators:

Prof. Dr. Arndt Bröder

Prof. Dr. Mandy Hütter

Contents

Summary

Non-threatening objects are more often misidentified as weapons when people are presented beforehand with Black compared to White male faces. This effect of race on object identification is well-established and has been reliably replicated using the Weapon Identification Task (WIT). The WIT is a sequential priming paradigm which instructs participants to identify target objects (i.e., guns vs. tools) after the presentation of face primes varying by race (i.e., Black vs. White males). However, the cognitive processes and mechanisms leading to weapon identification bias have been a matter of debate.

To further elucidate how this racially biased behavior is generated, this dissertation examines in three original research articles, respectively, the mechanisms leading to racial bias, the interplay of automatic and controlled processes in weapon identification, and the correspondence of different task procedures used to assess this effect. Manuscript 1 revealed that racial bias is mainly driven by response bias varying by race, meaning a stronger tendency to respond with "gun" after Black compared to White male faces. However, if participants engage in racial profiling, target discrimination is additionally biased by race. Manuscript 2 compared different process models which differ in their assumptions about the nature and temporal interplay of automatic and controlled processes in task performance. The Default Interventionist Model (DIM) emerged as the preferred model. The DIM posits an automatic initial default response which then may or may not be overcome by subsequent target discrimination and conflict resolution processes. Manuscript 3 investigated the correspondence of three implicit measures configured to assess the association of Black males with guns: the WIT, the First-Person Shooter Task (FPST), and the Implicit Association Test (IAT). All three measures were able to assess racial bias. The WIT and FPST displayed overall moderate correspondence in racial bias estimates indicating similarity in assessed construct and task procedure. In contrast, the IAT displayed mixed correspondence with the other two measures. The latter result may be explained by procedural specificities of the IAT such as race category salience and dual-categorization. Taken together, the findings of the three manuscripts help to get a better understanding of the complex interplay of cognitive mechanisms and processes leading to racial bias in weapon identification.

Manuscripts

This dissertation investigates cognitive processes underlying the Weapon Identification Task (WIT) by analyzing the mechanisms how race influences responding (Manuscript 1), the interplay of automatic and controlled processes (Manuscript 2), and the correspondence to other implicit measures (Manuscript 3). Manuscript 1 is submitted for publication, Manuscript 2 is published in *Social Cognition*, and Manuscript 3 is submitted for publication.

The research conducted in this dissertation was supported by the Research Training Group "Statistical Modeling in Psychology", funded by the Deutsche Forschungsgemeinschaft (GRK 2277), by a Doctoral Research Fellowship granted by the Foundation of German Business (Stiftung der Deutschen Wirtschaft, sdw gGmbH), funded by the German Federal Ministry of Education and Research, and by a one-year research grant for doctoral candidates funded by the German Academic Exchange Service (DAAD).

Manuscript 1

Laukenmann, R., & Erdfelder, E. (2023). *The Nature of Racial Bias in the Weapon Identification Task: Discrimination Bias, Response Bias, or Both?* Manuscript submitted for publication.

Manuscript 2

Laukenmann, R., Erdfelder, E., Heck, D. W., & Moshagen, M. (2023). Cognitive Processes underlying the Weapon Identification Task: A Comparison of Models accounting for Both Response Frequencies and Response Times. *Social Cognition*, *41*(2), 137–164. https://doi.org/10.1521/soco.2023.41.2.137

Manuscript 3

Laukenmann, R., & Calanchini, J. (2023). *Towards a Process-Level Understanding of Correspondence among Implicit Measures of Racial Bias.* Manuscript submitted for publication.

1. Introduction

Police shootings of unarmed Black men have resulted in more violent deaths compared to other racial groups (Kahn & Martin, 2020). Motivated by this disproportionate use of lethal force psychologists have investigated the processes leading to racially biased behavior of misidentifying harmless objects as weapons and consequently the decision to shoot an unarmed suspect (Correll et al., 2015; Klauer & Voss, 2008; Klauer et al., 2015; Payne, 2001; Payne & Correll, 2020). In consequence, psychologists are interested in how cultural stereotypes (i.e., the shared knowledge and assumptions about groups) influence behavior even when they exist outside of people's awareness and are not intentionally relied upon (Fish & Syed, 2020; Gawronski, 2019; Payne et al., 2017). More specifically, how do stereotypes of Black males seen as more aggressive and threatening than non-Black males result in racially biased behavior?

Implicit measures[1] are a way to assess the influence of racial stereotypes embedded in cultural knowledge. The main idea of these measures is that they capture information about psychological attributes (e.g., evaluations, stereotypes) without directly asking for them (Gawronski et al., 2020). Participants are typically instructed to identify or categorize stimuli (e.g., words or pictures), but are meanwhile exposed to additional stimuli containing race information. This may interfere with task performance and lead to biased response behavior. For example, in Payne's (2001) Weapon Identification Task, participants are instructed to identify guns and tools as fast and as correctly as possible, while race information is presented by a preceding face prime (i.e., Black vs. White males). While this task could reliably show that people tend to misidentify non-threatening objects more often as guns when paired with Black faces (Rivers, 2017), the cognitive processes and mechanisms leading to this behavior have been a matter of debate (Conrey et al., 2005; Klauer & Voss, 2008; Klauer et al., 2015; Payne, Shimizu, & Jacoby, 2005; Todd et al., 2021). For example: Are stereotypes interfering in response execution (Gawronski et al., 2010; Payne, Shimizu, & Jacoby, 2005) or are objects actually misperceived due to racial context cues (Correll et al., 2015; Klauer et al., 2015)? Furthermore, are stereotype associations influencing behavior all the time or only on some occasions (Klauer & Voss, 2008)?

The goal of this thesis is to get a better understanding of the cognitive processes and mechanisms that lead to racially biased behavior in misidentifying weapons and non-threatening objects. I will first present the experimental paradigm of the Weapon Identification Task (Payne, 2001; Rivers, 2017) and provide a brief overview of research conducted on factors influencing racial bias. I will then discuss cognitive processes proposed to influence task performance and then explicate the methodological

[1] I refer with the term "implicit measure" to a family of measurement instruments which assess information about psychological attributes (e.g., evaluations, stereotypes) without directly asking participants for that information (for an overview see Gawronski et al. 2020)

approach of the Process Dissociation Procedure (PDP; Jacoby, 1991; Payne, 2001) which has been predominantly used to analyze these processes. Subsequently, I provide an overview of the main findings of the three manuscripts. Finally, I discuss strengths and limitations of this line of research as well as future directions to investigate cognitive processes in weapon identification.

2. Weapon Identification and Racial Bias

2.1. The Standard Weapon Identification Task Paradigm

To assess the influence of racial primes on the perceptual identification of weapons, Payne (2001) proposed the Weapon Identification Task (WIT) as a sequential priming paradigm. Figure 1 shows a schematic illustration of a critical trial in the WIT. After an initial fixation cross in the center of the screen, participants are presented with a race prime (i.e., a face of a Black or White male) followed by a target object (i.e., a depiction of a gun or tool) which is then covered by a pattern mask. Participants are instructed to identify the target object as "gun" or "tool" as accurately and quickly as possible.

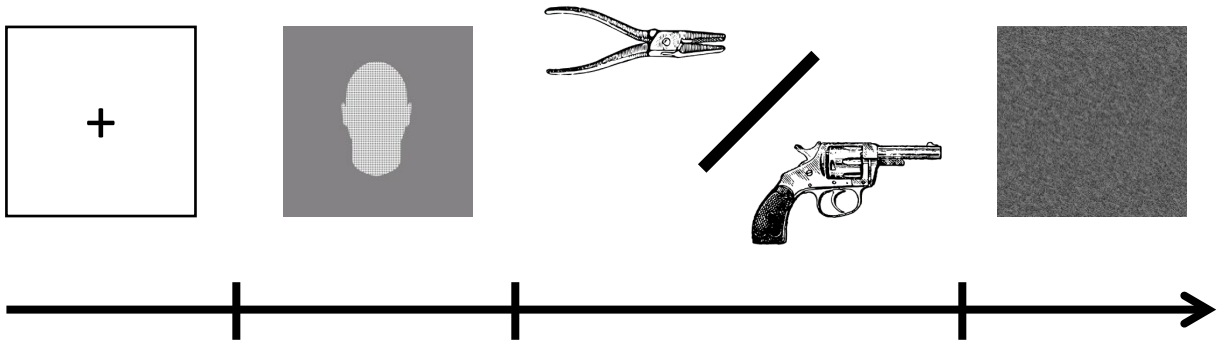


Figure 1. Schematic illustration of the temporal order of a critical trial in a standard Weapon Identification Task (WIT). Face primes vary by race (typically a Black or a White male face) and target objects by object category (typically a tool or a gun object, here taken from Rivers, 2017) from trial to trial. The face prime is represented here by a neutral face outline. Manuscripts 1 to 3 analyzed variations of the standard paradigm of the WIT.

One of the main variations in implementing the WIT is the response time window provided for participants (Payne, 2001; Rivers, 2017). A short response time window typically leads to a difference in response accuracy and a long response time window typically leads to differences in response times for correct object identification. Regarding correct response times, participants identify gun targets faster if preceded by Black faces compared to White faces. Regarding error rates, participants misidentify tool targets as guns more often if preceded by Black faces compared to White faces. These result patterns, reflected in response times and error rates, are known as weapon identification bias (Payne, 2001; Rivers, 2017; Todd et al., 2021).

A meta-analysis for the WIT by Rivers (2017) reports a large effect size for error rates (η_p^2 = .204; 95%-CI = [.151 – .266]; N_{study} = 33) and a medium effect size for correct response times (η_p^2 = .106; 95%-CI = [.039 – .208]; N_{study} = 15), both with likely evidential value and unlikely being biased

by extensive *p*-hacking according to *p*-curve analysis. Similarly, a meta-analysis of sequential stereotype priming tasks by Kidder et al. (2018) reports for the WIT a small to medium effect size ($d_Z = 0.46$; 95%-CI = [0.36 – 0.56]; $N_{study} = 30$) for error rates and correct response times combined. Hence, the WIT is able to assess an identification bias for weapons by race.

2.2. Factors Affecting Weapon Identification Bias

Although the WIT shows a weapon identification bias effect across several studies, its size is moderated by a multitude of factors. More specifically, identification bias is influenced by prime characteristics, task instructions, experimental manipulations, and individual differences in participants.

Regarding prime characteristics, the conventional version of the WIT uses Black and White adult male faces with emotionally neutral expression as primes (e.g., Amon & Holden, 2016; Correll, 2008; Klauer et al., 2015; Klauer & Voss, 2008; Lambert et al., 2003; Madurski & LeBel, 2015; Payne, 2001; Payne et al., 2002; Rivers, 2017). However, several other prime characteristics may lower or increase weapon identification bias. For example, the emotional expression of face primes, balanced by prime race, modulates identification bias. Additionally angry Black male faces elicit a more pronounced weapon identification bias compared to happy Black male faces, which do not elicit any effect (Kubota & Ito, 2014). Furthermore, weapon identification bias may vary by other social dimensions of the primes, like age and gender. Concerning age, racial weapon identification bias is robust and generalizes across age ranging from young Black boys (Todd et al., 2016) to elderly Black men (Lundberg et al., 2018). Thus, regardless of age, Black males were associated more strongly with guns than White males. In addition, regardless of race, young boys were less associated with guns than adults, whereas young adults and elderly men were equally strongly associated with guns (Lundberg et al., 2018; Todd et al., 2016). Concerning gender, males are more associated with guns than females, although this effect is qualified by race as the gender difference only emerges for Black but not for White faces (Thiem et al., 2019). Overall, prime characteristics like emotional expression and social dimensions can modulate weapon identification bias.

Task instruction and category salience also play a role in WIT performance. Typically, participants are instructed to disregard the face primes for object identification (Payne, 2001; Payne et al., 2002; Rivers, 2017), but if they are instructed to engage in racial profiling or to explicitly avoid the usage of the primes' race, weapon identification bias is more pronounced (Payne et al., 2002). On the other hand, if participants use implementation intentions (i.e., predefined if-then action plans) to think "safe" when they see a Black face, weapon identification bias is attenuated (Stewart & Payne, 2008). Similarly, category salience modulates weapon identification bias. In a WIT including Black and White boys and adults, category salience, induced by a beforehand performed sorting task, led to a stronger racial bias in weapon identification when race was made salient in comparison to age (Todd et al., 2021).

Hence, variations in instructions and category salience can modulate the effect size of weapon identification bias in both directions.

In addition to instructions, experimental manipulations can attenuate the overall expression of errors by influencing cognitive resources and control participants are able to dedicate to the task. For example, a shorter response time window (Payne, 2001; Payne et al., 2002), cognitive depletion induced by a preceding Stroop task (Govorun & Payne, 2006), and social anxiety induced by announced public discussion of task performance (Lambert et al., 2003), all diminish the amount of controlled responding in the WIT. This indicates that participants need replenished cognitive resources to reach a high number of correct responses.

In a similar vein, individual differences in cognitive abilities and motivations are connected to the performance in the WIT. Ito et al. (2015) report that controlled processes leading to correct responses correlate positively with participants' cognitive executive function abilities entailing response inhibition, for example. On a motivational account, participants' internal motivation to control prejudice tends to correlate with controlled responding, whereas external motivation to control prejudice does not (Ito et al., 2015; Volpert-Esmond et al., 2020). Hence, individual executive function capabilities and motivation to control prejudice can modulate the expression of weapon identification bias.

Overall, this variability of weapon identification bias and correct responding caused by different factors suggests that different cognitive processes play a role in the performance of the WIT. Different prime characteristics (e.g., emotional expression, age, gender) and salience of feature dimensions (e.g., category salience) lead to a variation in strength of weapon identification bias, which is tied to the stereotype associations elicited by automatic processes. Whereas participants' cognitive resources and capabilities (e.g., cognitive depletion level, response inhibition abilities, motivation) relate to overall task performance in correct responding, which is tied to executive control processes. In the General Discussion, I will come back and discuss how these factors relate to the insights gained by the three manuscripts.

3. Cognitive Processes in Weapon Identification

3.1. Cognitive Processes Underlying Implicit Measures

Implicit measures[2], like the WIT, were designed to assess the strength of mental associations[3] between target groups (i.e., Black and White males) and their ascribed group features (i.e., cultural stereotypes), while aiming to minimize the interference of other processes, such as socially desirable responding. However, implicit measures are not process pure and reflect other processes besides mental associations (Calanchini et al., 2014; Gawronski, 2019; Payne, 2001).

For the WIT, two types of processes are typically discussed (Payne, 2001; Klauer & Voss, 2008): controlled and automatic processes. Controlled processes represent stimulus discrimination efforts that rely on cognitive resources and aim at an accurate representation of the target stimulus (i.e., gun or tool targets). Automatic processes represent simple associations or habitual responses assumed to be effortless, spontaneous, and triggered by the environment (e.g., Black or White face primes) or individual preferences (e.g., handedness). Across several studies, racial bias in the WIT was associated with automatic processes which trigger a stronger threat stereotype association of Black males compared to White males, whereas controlled processes were not influenced by race (Huntinsger et al., 2008; Ito et al., 2015; Klauer & Voss, 2008; Lundberg et al., 2018; Payne, 2001; Payne et al., 2002; Todd et al., 2016; Thiem et al., 2019; but see Klauer et al., 2015). A way to estimate the influence of these different cognitive processes is using formal mathematical models that disentangle the contribution of different processes in task performance (Gawronski, 2019; Sherman et al., 2010). One of the most prominent models used to analyze the WIT is the Process Dissociation Procedure (PDP; Jacoby, 1991; Payne, 2001).

[2] The term "implicit" stimulated a debate in psychological literature due to an inconsistent use of the term. For discussions on the meaning of the term "implicit", see Corneille and Hütter (2020), and Gawronski and Brannon (2019). In this work, I use the term "implicit measure" in alignment with Gawronski et al., (2020, see also, Fazio & Olson, 2003) in reference to a family of measurement instruments.

[3] The nature of the mental representations in implicit measures having an associative or propositional structure are still a matter of debate (Brownstein et al., 2019; De Houwer, 2009; Mandelbaum, 2016). This thesis makes no claims about the nature of the mental representations. However, in alignment with previous research on the Weapon Identification Task, these mental representations are referred to as associations.

3.2. The Process Dissociation Procedure

The PDP is a formal mathematical model which aims to disentangle the influence of two latent processes on task performance in the WIT, labeled controlled responding and automatic response bias (Klauer & Voss, 2008; Payne, 2001). The *C*-Parameter represents the probability of successful controlled processes, resulting in correct responses. In contrast, the *A*-parameter represents automatic response bias, which reflects participants' preference to respond with gun compared to tool. As a prerequisite, the PDP needs to be applied to an experimental paradigm which provides frequency data, like the WIT, which records participants' errors and correct responses for gun and tool targets.

Figure 2 illustrates the cognitive architecture of the PDP in a Multinomial Processing Tree (MPT; Riefer & Batchelder, 1988). An MPT depicts different cognitive pathways, or branches, which can lead to a target response. In an MPT model, parameters represent probabilities that a process drives the response in the one or the other direction. For the PDP, two process pathways can lead to the participants' response pattern. The upper panel of Figure 2 illustrates the process pathways for gun targets. In the first pathway, the controlled process succeeds (C), which always leads to a correct target discrimination, hence a correct response. In the second pathway, the controlled process fails to discriminate the target correctly ($1 - C$), but the participants' response bias towards guns (A) gives the correct response. Conversely, if the controlled process fails ($1 - C$), participants' response bias towards tools ($1 - A$) gives the incorrect response. For a tool target (lower panel Figure 2), the cognitive pathway structure is the same, but, in contrast to the gun target condition, response bias towards guns (A) results in an incorrect response and response bias towards tools ($1 - A$) in a correct response.

To estimate the parameters C and A of the PDP using MPT modeling (for a tutorial, see Schmidt, et al., 2023), the observed response frequencies are equated with the expected response probabilities of the response categories. The expected response probabilities consist of the sum of branch probabilities resulting in the same response. The probability of a branch consists of the product of probability parameters of the respective branch. For example, the expected response probability of a gun response for a gun target trial consists of the sum of the branch for a successful controlled process (C) and the branch for a response bias towards gun if the controlled process fails ($(1 - C) \cdot A$). This results in the model equation:

$$p(\text{gun response} \mid \text{gun target}) = C + (1 - C) \cdot A \tag{1}$$

An MPT model then formalizes a system of model equations for all response categories for every condition. Hereupon, model parameters can be estimated based on the response frequency pattern of a given dataset (Heck, Arnold, & Arnold, 2018; Klauer, 2010; Moshagen, 2010; Schmidt et al., 2023).

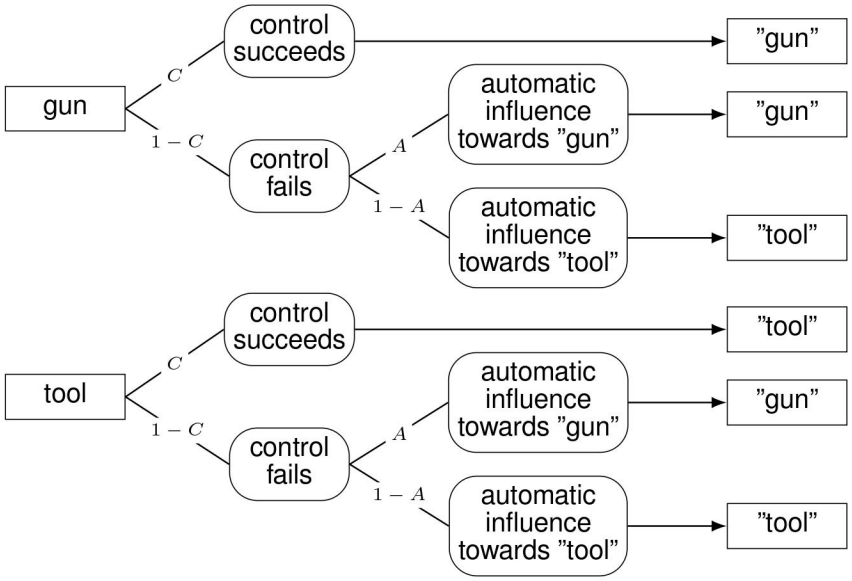


Figure 1. Process Dissociation Procedure (PDP) for the Weapon Identification Task (Figure taken from Laukenmann et al., 2023). Parameters C and A denote probabilities of response determination by a controlled process and an automatic process, respectively. Note that A is conditional on a failure of the controlled process, that is, A represents the conditional probability of response determination by an automatic process given controlled process failure.

The PDP has often been used to estimate the influence of primes' race on controlled processes (C), which represent successful target discrimination, and automatic processes (A) which represent the extent of response bias. Typically, studies on the WIT report that controlled process parameters do not vary between Black and White prime faces, but that response bias towards guns is larger following a Black than a White face prime (Huntinsger et al., 2008; Ito et al., 2015; Lundberg et al., 2018; Payne, 2001; Payne et al., 2002; Todd et al., 2016; Thiem et al., 2019; but see Klauer et al., 2015). This pattern is interpreted as the result of the influence of automatic stereotype associations elicited by the primes' race. Black males are associated more strongly with danger, and in consequence associated more strongly with the gun target. Thus, participants show an increased response bias to respond with gun following Black compared to White face primes.

3.3. Investigating the Nature of Cognitive Processes in the WIT

The PDP has been widely used to analyze the influence of different cognitive processes in the WIT (e.g., Ito et al., 2015; Klauer et al., 2015; Payne, 2001). Nevertheless, different aspects regarding the interpretation of the process parameters need more thorough investigation. That is why, this dissertation

investigates the processes underlying the Weapon Identification Task using the Process Dissociation Procedure as reference framework.

Manuscript 1 investigates whether the influence of the primes' race on participants' performance is mediated by target discrimination or response bias (i.e., *C*- and *A*-parameter of the PDP). This is tested for participants performing the standard version of the WIT and when they were explicitly instructed to engage in racial profiling. Furthermore, Manuscript 1 looks at a basic model assumption of the conventionally applied version for the PDP. Specifically, it investigated whether the *C*-parameter in the PDP can be equated across target object conditions or not (Klauer et al., 2015).

Manuscript 2 investigates the nature and temporal interplay of the processes the PDP-parameters are assumed to represent. The *A*-parameter can reflect several automatic processes like the activation of racial associations or a guessing tendency. The *C*-parameter can reflect several controlled processes like perceptual target discrimination or conflict resolution if racial associations interfere with target identification. Based on that, different psychological process models for the PDP are conceivable, which vary by their nature and temporal order they assume for the *C*- and *A*-parameter.

Manuscript 3 investigates the correspondence of the WIT with other implicit measures of racial bias, which were developed to measure the association of Black males with weapons and threat: the First-Person Shooter Task (FPST; Correll et al., 2002) and the Implicit Association Test (IAT; Greenwald et al., 1998) using weapons and tools as stimuli. Beyond comparing task performance directly, Manuscript 3 compares different cognitive process models which may be qualified by the different procedural setups of these tasks. Hence, the correspondence between parameters reflecting the influence of controlled or automatic processes in task performance, can be compared relying on the respective appropriate process model for these tasks.

In the following, I will outline the substantive focus of each manuscript, its methodological approach, and the main results. The manuscripts will provide a better understanding about the cognitive processes underlying the performance in the Weapon Identification Task and their relation to other implicit measures.

4. Cognitive Processes Underlying the WIT

4.1. Manuscript 1: The Influence of Race on Target Discrimination and Response Bias

In the first manuscript, we investigated mechanisms leading to racially biased responding in the WIT. Two mechanisms have been proposed to drive racial bias (Klauer & Voss, 2008; Klauer et al., 2015; Payne, Shimizu, & Jacoby, 2005; Todd et al., 2021). The first mechanism is discrimination bias, meaning that race influences target discrimination by modifying the perception and information extraction from the target object. The second mechanism is response bias, meaning that primes elicit an additional stream of information of threat-stereotype-based associations which leads response bias to vary by race. Both explanations are plausible mechanisms for how prime race can affect participants' task performance, and they are not mutually exclusive.

If discrimination bias drives racial bias in the WIT, then prime race affects perception and interpretation of parts of the target objects. Specifically, prime race provides stereotype-biased context cues to resolve perceptual ambiguity in object discrimination. For example, a metal tube might be more easily interpreted as the barrel of a gun after Black male face primes and more easily as the shaft of a screwdriver after White male face primes (Klauer et al., 2015; Klauer & Voss, 2008; Payne, Shimizu, & Jacoby, 2005). If response bias varies by prime race, threat-stereotypes triggered by the primes' race elicit an additional stream of information, alongside the information from the target object, resulting in a stronger preference to respond with "gun" after Black faces and with "tool" after White faces. Previous studies found support for racial bias mediated by response bias (Klauer & Voss, 2008; Payne, Shimizu, & Jacoby, 2005; Todd et al., 2021), whereas evidence for discrimination bias is mixed. Some studies report support for this mechanism (Klauer et al., 2015) but others do not (Payne, Shimizu, & Jacoby, 2005; Todd et al., 2021).

A feasible way to investigate these mechanisms is using the PDP as both mechanisms map on different process parameters (Klauer et al., 2015). While discrimination bias is represented by controlled processes mapping on the *C*-parameter, response bias is represented by automatic processes mapping on the *A*-parameter. More specifically, if threat-stereotype associations influence response bias, this should lead to a higher preference to respond with "gun" after Black face primes (*B*) compared to White face primes (*W*), hence in a larger *A*-parameter in the Black compared to the White face prime condition (i.e., $A_B > A_W$). If race biases target discrimination, the *C*-parameter should vary across prime and target conditions. Specifically, Black face primes should enhance the target discrimination for guns (*G*) compared to tools (*T*), resulting in a larger *C*-parameter in the gun compared to the tool target condition (i.e., $C_{BG} > C_{BT}$). In contrast, White face primes should enhance the target discrimination for tools (*T*) compared to guns (*G*), resulting in a larger *C*-parameter in the tool compared to the gun target condition

(i.e., $C_{WG} < C_{WT}$). Please note that the conventional version of the PDP applied to the WIT typically estimates only four parameters (i.e., C_B, C_W, A_B, and A_W), hence the C-parameter estimates are normally restricted across target object conditions. To develop a generalized PDP variant which allows the C-parameter to vary across prime and target conditions, one can gain more degrees of freedom for modeling by including an additional within-subjects manipulation that selectively affects response bias but leaves target discrimination unaffected (Klauer et al., 2015).

We conducted two studies in Manuscript 1, testing whether racial bias in the WIT is mediated by target discrimination, response bias, or both. In both studies, we implemented an additional within-subjects manipulation which consisted of a base-rate manipulation for target objects. This results in two experimental blocks, one with more gun than tool targets (70%:30%) and the other vice versa (30%:70%). The first study applied a standard WIT instructing participants to identify gun and tool targets while disregarding the prime faces. The second study applied a standard WIT as a control group and an experimental group explicitly instructed to rely on racial profiling in weapon identification. We used the R package TreeBUGS (Heck et al., 2018) to estimate the generalized PDP as a hierarchical latent-trait MPT model. Parameter estimates for the generalized PDP are listed in Table 1.

Table 1. Parameter estimates for generalized PDP model of Study 1 and 2 of Manuscript 1

	Study 1		Study 2			
			Control group		Racial profiling group	
Parameter	Mean	95%-BCI	Mean	95%-BCI	Mean	95%-BCI
$C_{B,G}$	.306	[.255 – .358]	.381	[.329 – .435]	.402	[.347 – .458]
$C_{B,T}$	.202	[.149 – .257]	.279	[.223 – .335]	.236	[.180 – .294]
$C_{W,G}$	.320	[.273 – .366]	.365	[.315 – .417]	.284	[.233 – .332]
$C_{W,T}$	.221	[.166 – .273]	.288	[.229 – .344]	.309	[.245 – .369]
$C_{B,G}/C_{B,T}$	1.544	[1.167 – 2.078]	1.377	[1.109 – 1.735]	1.725	[1.348 – 2.233]
$C_{W,G}/C_{W,T}$	1.470	[1.127 – 1.956]	1.278	[1.047 – 1.576]	0.928	[0.741 – 1.144]
C_{IA}	1.064	[0.752 – 1.463]	1.085	[0.830 – 1.384]	1.877	[1.373 – 2.549]
$A_{B,MG}$	.766	[.743 – .789]	.770	[.745 – .794]	.784	[.755 – .810]
$A_{W,MG}$	.723	[.696 – .749]	.750	[.723 – .776]	.728	[.699 – .756]
$A_{B,MT}$	.217	[.197 – .239]	.230	[.207 – .254]	.271	[.242 – .300]
$A_{W,MT}$	.183	[.164 – .203]	.180	[.162 – .200]	.201	[.179 – .223]
$A_{B,MG}/A_{W,MG}$	1.059	[1.020 – 1.103]	1.027	[0.993 – 1.063]	1.078	[1.027 – 1.128]
$A_{B,MT}/A_{W,MT}$	1.187	[1.061 – 1.326]	1.280	[1.141 – 1.432]	1.352	[1.155 – 1.564]

Note. C = controlled process, A = automatic process, B = Black prime, W = White prime, G = Gun target, T = Tool target, MG = More Guns base-rate condition, MT = More Tools base-rate condition. $C_{IA} = (C_{B,G}/C_{B,T})/(C_{W,G}/C_{W,T})$. BCI = Bayesian Credibility Interval.

Across both studies, results indicated an effect on response bias varying by race. The ratio of the *A*-parameters by race (i.e., A_B/A_W) were reliably larger than one except for the more guns base-rate condition of the control group in Study 2. However, no interaction of the *C*-parameter by prime and target conditions emerged for Study 1 and the control group of Study 2. This interaction is represented by taking the ratio of the *C*-parameter for gun and tool targets in the Black prime condition divided by its respective counterpart in the White prime condition (i.e., $C_{IA} = (C_{B,G}/C_{B,T})/(C_{W,G}/C_{W,T})$). However, for the racial profiling group in Study 2, this interaction was reliably greater than one, which indicates that prime race biases target discrimination. To reiterate, race moderated response bias across studies, whereas discrimination bias only emerged when participants were explicitly instructed to use the primes' race for target identification.

Our findings align well with previous findings, reporting response bias as a major mechanism driving racially biased responding in the WIT (Klauer & Voss, 2008; Todd et al., 2021). This means that primes' race automatically elicits threat-stereotype associations creating a second stream of information providing a response alternative besides information extracted from the target object. Hence, racially biased responses in the WIT emerge when target discrimination was not successful (Payne, Shimizu, & Jacoby, 2005; Klauer & Voss, 2008). Furthermore, our findings suggest that primes' race does not influence target discrimination in the standard version of the WIT, but it does if participants engage in racial profiling. This can explain the diverging findings in literature as some authors reported racially biased target discrimination (Klauer et al., 2015) but others did not (Todd et al., 2021). For example, participants might spontaneously engage in racial profiling as response strategy even if researchers had not instructed them to do so.

From a modeling perspective, Manuscript 1 also investigated the assumption of the conventionally used PDP to restrict the *C*-parameter across target object conditions, which results in one controlled process parameter for Black primes and another for White primes (i.e., C_B, and C_W). This is the most widely used specification of the PDP when applied to the WIT (e.g., Huntsinger et al., 2009; Ito et al., 2015; Payne, 2001; Payne et al., 2002; Thiem et al., 2019). In our studies, the *C*-parameter did not vary as a function of target and prime in the standard version of the WIT. This lends support for the conventional PDP specification to restrict the *C*-parameter across target object conditions. However, this assumption may be violated if participants rely on primes' race as context cue for object identification, because the *C*-parameter varied as a function of target and prime in the racial profiling condition. Hence, researchers should take into account whether possible experimental manipulations might lead to violations in the assumptions of the conventionally used parameter restrictions for the PDP.

Complementing the main research question in Manuscript 1, these findings also inform whether an alternative model specification to the conventional PDP is advisable if only the standard four

conditions of the WIT are available (i.e., Black versus White primes crossed with guns versus tools). In addition to the conventionally used form of the PDP, which estimates the four parameters (i.e., C_B, C_W, A_B, and A_W) an alternative model specification is conceivable. Specifically, this alternative PDP model restricts C-parameters across primes resulting in two separate C-parameters for guns and tools next to two separate A-parameters for Black and White primes (i.e., C_G, C_T, A_B, and A_W). Consequently, this model allows for a main effect of the target object on controlled responding instead of prime race. This main effect of target object on controlled responding was observed in Study 1 and the control group of Study 2. Participants had larger C-parameters for gun trials compared to tool trials whereas prime race showed no main effect on the C-parameters. Thus, the current evidence best supports the alternative PDP (with C_G, C_T, A_B, and A_W) compared to the conventional PDP.

In addition, we like to point out, that the specification of the C-parameters in the PDP influences the estimates for A_B and A_W. This is crucial, as for example if the alternative PDP is the true model for the standard WIT paradigm, the estimates for A_B and A_W of the conventional PDP might be biased. This can be demonstrated by plugging in the expected error rate probabilities resulting from the alternative PDP model in the equations for the conventional PDP model. The model equations for error rates in the conventional PDP are:

$$p(\text{tool response} \mid \text{Black prime, gun target}) = (1 - C_B) \cdot (1 - A_{B,conventional}) \quad (2)$$

$$p(\text{gun response} \mid \text{Black prime, tool target}) = (1 - C_B) \cdot A_{B,conventional} \quad (3)$$

$$p(\text{tool response} \mid \text{White prime, gun target}) = (1 - C_W) \cdot (1 - A_{W,conventional}) \quad (4)$$

$$p(\text{gun response} \mid \text{White prime, tool target}) = (1 - C_W) \cdot A_{W,conventional} \quad (5)$$

The model equations for error rates in the alternative PDP are:

$$p(\text{tool response} \mid \text{Black prime, gun target}) = (1 - C_G) \cdot (1 - A_{B,alternative}) \quad (6)$$

$$p(\text{gun response} \mid \text{Black prime, tool target}) = (1 - C_T) \cdot A_{B,alternative} \quad (7)$$

$$p(\text{tool response} \mid \text{White prime, gun target}) = (1 - C_G) \cdot (1 - A_{W,alternative}) \quad (8)$$

$$p(\text{gun response} \mid \text{White prime, tool target}) = (1 - C_T) \cdot A_{W,alternative} \quad (9)$$

So, to demonstrate how $A_{B,conventional}$ relates to $A_{B,alternative}$, one can equate the ratio of errors on tool to gun targets between the different model specifications (i.e., Equation (2)/Equation (3) = Equation (6)/Equation (7)) and solve it by $A_{B,conventional}$. This results in the following equation:

$$A_{B,conventional} = ((1 - C_T) \cdot A_{B,alternative})/(((1 - C_G) \cdot (1 - A_{B,alternative})) - ((1 - C_T) \cdot A_{B,alternative})) \quad (10)$$

The same equation is true for the A_W-parameters as Equations (4), (5), (8), and (9) are structurally analogous to Equations (2), (3), (6), and (7). Overall Equation (10) reveals that $A_{B,conventional}$ is equal to $A_{B,alternative}$ when the C-parameter does not vary by target object (i.e., C_G =

C_T). But if the C-parameter varies by target object (i.e., $C_G \neq C_T$) the estimate for $A_{B,conventional}$ is skewed. In specific, if the C-parameter for gun targets is smaller than for tool targets (i.e., $C_G < C_T$), it leads to a smaller estimate of $A_{B,conventional}$ compared to the initial value of $A_{B,alternative}$. In contrast, if the C-parameter for gun targets is larger than for tool targets (i.e., $C_G > C_T$), it leads to a larger estimate of $A_{B,conventional}$ compared to the initial value of $A_{B,alternative}$. For example, assuming a larger C-parameter for gun than tool targets $C_G = .30$ and $C_T = .20$, as found in Study 1 and the control group of Study 2, and assuming typically observed A-parameter values $A_{B,alternative} = .60$ and $A_{W,alternative} = .50$ (Ito et al., 2015; Klauer et al., 2015; Payne, 2001) for the alternative PDP, this results in larger A-parameter estimates in the conventional PDP: $A_{B,conventional} = .632$ and $A_{W,conventional} = .533$. However, the difference score between A_B and A_W is similar, but nevertheless slightly underestimated in the conventional PDP if $C_G > C_T$.

In sum, researchers should note that if controlled processes vary by target objects (e.g., higher controlled processing for gun targets), then the conventional PDP may result in biased A-parameter estimates. Thus, if applying the PDP to the standard paradigm of the WIT (Black versus White primes crossed with guns versus tools), the alternative PDP may provide a better representation of processes estimates than the conventional PDP.

Returning back to the overall conclusions of Manuscript 1, apart from PDP model specification, the first manuscript identified response bias as the main mechanism driving racial bias in the WIT, whereas discrimination bias only became important when participants used primes' race for object identification. Nevertheless, the specific nature and temporal interplay of the automatic and controlled processes in the WIT remain unspecified in Manuscript 1. We therefore investigated these aspects in Manuscript 2.

4.2 Manuscript 2: Psychological Process Models for the WIT

The goal of the second manuscript was to investigate the nature and interplay of the cognitive processes influencing task performance in the WIT. The PDP estimates the probabilities of controlled and automatic processes but remains agnostic about their psychological characteristics as well as their temporal sequence. Conrey et al. (2005) discussed four potential cognitive determinants leading to responses in the WIT: the general ability to discriminate gun and tool objects, the activation of racial stereotype associations, the ability to resolve response conflicts between racial associations and target identification (so-called overcoming bias), and the guessing tendency towards one of the response options. These four cognitive determinants closely relate to the controlled and automatic processes in the PDP. More precisely, target discrimination and response conflict resolution reflect controlled processes, whereas stereotype association activation and guessing tendency reflect automatic processes (Conrey et al., 2005; Klauer & Voss, 2008; Sherman, 2006). Moreover, these processes may run in parallel

or in sequence, as the PDP states that the automatic process only drives the response when the controlled process fails. Consequently, this reflects a conditional relation, which does not imply that automatic processes follow controlled processes in a temporal fashion. In fact, several psychological process models based on the PDP are conceivable, relying on different assumptions about the nature and interplay of the cognitive processes.

Based on Evans (2007), Klauer and Voss (2008) discussed four different psychological process models for the WIT which are all instantiations of the PDP: the Preemptive Conflict-Resolution Model (PCRM), the Default Interventionist Model (DIM), the Parallel Competitive Model (PCM), and the Guessing Model (GM). The PCRM assumes a preemptive decision before every trial whether either the controlled or the automatic process will determine the response. The DIM assumes that an automatic default response is always activated first providing a response which then may or may not be overcome by a conflict resolution process. The PCM assumes that automatic and controlled processes run in parallel, both providing a response, sometimes requiring a subsequent conflict resolution process if they provide diverging responses. The GM assumes that the controlled process determines the response, but if it fails to do so, the response is determined by guessing. Overall, these process models possess the same conditional process structure of the PDP based on the accuracy pattern of responses, but they differ in their assumptions for the relative latencies of process branches. Previous work, which relied on the comparison of overall response times for correct and false responses across conditions, indicated support for the PCRM and DIM but was unable to determine a final preference for one or the other (Klauer & Voss, 2008).

In Manuscript 2, we conducted a comparison of these four process models by reanalyzing previously published data sets of the WIT. To compare these process models, we formalized them in the framework of response time-extended Multinomial Processing Tree models (MPT-RT; Heck & Erdfelder, 2016). This allows us to jointly estimate the core parameters C and A of the PDP as well as parameters representing the relative latency of each processing branch. In addition, the MPT-RT approach can test different sets of equality constraints and order restrictions on the parameters (Knapp & Batchelder, 2004) as those implied by the four process models. For our model comparison, we reanalyzed eight data sets of the WIT in its conventionally used version which relies on Black and White male faces as primes without any other experimental manipulation.

In our first model comparison we specified each process model as MPT-RT with order restrictions implemented on latency parameters respective to each models' assumptions. Results from the first comparison were clear-cut: the models for the DIM and PCRM provided acceptable model fit. In contrast, the models for the PCRM and GM provided blatant model misfit across all data sets. Both models, the DIM and PCRM, assume fast automatic and slow controlled process branches. Hence, the reason for their good model fit is that they both are able to predict fast error and slow correct response

latencies as typically observed in the WIT. The models for PCM and GM have difficulty accommodating this response pattern. Overall, the results of the first model comparison replicated the findings of Klauer and Voss (2008), showing a preference for the DIM and PCRM as process models underlying the WIT.

Our second model comparison investigated additional assumptions by restricting latency parameters across prime and target conditions. This provides more degrees of freedom for model estimation. These restrictions were later implemented in the third model comparison conducted to differentiate the DIM and PCRM. To be specific, our second model comparison investigated whether automatic process branches are equally fast independent of prime race and target object condition. This assumption is plausible as automatic associations are instantly triggered by the prime and thus are independent of the following target object. Additionally, even if relative response latencies of automatic process branches slightly differ, this difference should be negligible compared to the relatively slow controlled process branch latencies. Regarding controlled process branch latencies, we investigated whether controlled process branches differ by target type but not prime race. This assumption is plausible as participants tend to show faster correct responses for gun than tool targets (Payne, 2001). The difference may be driven by a threat-superiority effect leading to a faster identification of gun targets (Rivera-Rodriguez et al., 2021; Subra et al., 2018). Furthermore, this assumption is in accordance with the previous finding in Manuscript 1 that target discrimination is not affected by prime race for the standard version of the WIT. Results of the second model comparison showed that these restrictions on latency parameters of automatic and controlled processes resulted in models with acceptable model fit across almost all data sets.

The goal of the third model comparison was to determine whether the PCRM or the DIM provides better model fit for the WIT while relying on the additional assumptions tested in the second model comparison. For comparing the PCRM and DIM, we implemented an additional latency order restriction for the DIM. Specifically, the DIM assumes that an automatic default response is elicited at the beginning of each trial which then may or may not be overcome by a conflict resolution process. In consequence, there exist two types of controlled process branches: one branch where automatic default response and the to-be-identified target are congruent and the other branch where automatic default response and the to-be-identified target are incongruent. Hence, the later branch needs additional process time for resolving the response conflict. This leads to three process branches with different relative latencies: the fastest for automatic process branches, intermediate for controlled process branches with a congruent default response and the slowest for controlled process branches with an incongruent default response. In contrast, the PCRM assumes that controlled process branches do not vary in processing time. Because a preemptive decision determines whether the response is given by either the relatively slower, controlled process branches or the faster, automatic process branches. Results of the third model

comparison showed that both models fit well to nearly all data sets. But additionally, the DIM was preferred over the PCRM for the majority of data sets.

The preference for the DIM as a process model for the WIT aligns well with previous findings. The DIM posits that automatic stereotype associations are elicited at the onset of every trial followed by controlled processes. These controlled processes entail target identification and response conflict resolution if the default response and the to-be-identified target are incongruent. The DIM can explain the persistence of racial bias when participants' overall accuracy is high, and when weapon identification bias is reflected in correct response times (Klauer & Voss, 2008). For example, a faster response to gun targets after a Black male face may be the result of less cognitive conflict in comparison to tool targets. Additional support comes from research on brain region activation after seeing Black and White faces (Cunningham et al., 2004). Specifically, short presentation of faces led to a higher amygdala activation for Black than White faces. This activation of the amygdala is associated with higher emotionality and perceived threat. However, a longer presentation of faces revealed no difference in amygdala activation but increased activity for the dorsolateral prefrontal cortex and the anterior cingulate. Both are associated with controlled processing. In line with the DIM, this suggests that at first emotional reaction and threat associations are instantly activated but later controlled processes can inhibit and modulate these associations. By implication, cognitive capabilities may modulate the expression of weapon identification bias as target identification and conflict resolution require cognitive resources. As mentioned before, a shorter response time window (Payne, 2001; Payne et al., 2002), cognitive depletion (Govorun & Payne, 2006), and social anxiety (Lambert et al., 2003) led to less controlled responding. In addition, the important role of conflict resolution capabilities is substantiated by the fact that controlled processing in the WIT correlates with participants' general cognitive executive function abilities, for example, response inhibition abilities (Ito et al., 2015) and the internal motivation to control prejudice (Volpert-Esmond et al, 2020).

In conclusion, Manuscript 2 identified the Default Interventionist Model as the preferred model to describe the interplay of cognitive processes in the Weapon Identification Task. This implies that automatic process branches are faster than controlled process branches. Primes elicit an automatic default response moderated by racial threat-stereotype from the outset of each trial. In consequence, besides target object identification, response conflict resolution processes play a vital role in task performance of the WIT.

4.3. Manuscript 3: Correspondence of Cognitive Processes Underlying Different Implicit Measures of Racial Bias

Manuscript 1 and 2 took a thorough look at the mechanisms driving racial bias in the WIT as well as how underlying cognitive processes interact. Manuscript 3 broadens the perspective and investigates

how the WIT relates to other implicit measures of racial threat-stereotypes. More specifically, Manuscript 3 addressed the process-level correspondence of the WIT, the First-Person Shooter Task (FPST; Correll et al., 2002), and the Implicit Association Test (IAT; Greenwald et al., 1998). This is done by applying Multinomial Processing Tree models tailored to the three tasks to the same sample of participants.

Previous work on correspondence among different implicit measures revealed rather weak or even zero correlations between different implicit measures of racial bias (Bar-Anan & Nosek, 2014; Cunningham et al., 2001; Glaser & Knowles, 2008; Ito et al., 2015; Olson & Fazio, 2003; Payne, 2005). This is surprising for measures designed to assess a common construct. However, even when implicit measures are configured to assess a common construct, a wide variety of procedural differences among measures remains such as stimulus materials, response time limits, number of trials, task instructions, and stimulus presentation procedures. These differences in task procedures may determine which processes influence the performance in implicit measures (Gawronski et al., 2010). Consequently, procedural differences may result in weak correlations between implicit measures configured to assess a common construct.

In Manuscript 3, we aligned the three measures: WIT, FPST, and IAT across several procedural dimensions (e.g., using the same stimulus materials, imposing the same response time limit, and implementing the same number of trials). Nevertheless, other procedural aspects are unique to each of these tasks, for example, instructions and stimulus presentation procedures as detailed in the following. The WIT is a priming paradigm displaying the target object (e.g., guns vs. tools) after a preceding target group prime (e.g., Black vs. White male faces). Hence, primes and targets are presented sequentially. Furthermore, in the standard version of the WIT, participants are told to identify the target object while disregarding the face prime (Payne, 2001). The FPST is a search task displaying a member of the target group as a full body image (i.e., Black vs. White males) either armed (e.g., holding a gun) or unarmed (e.g., holding a cell phone). Target subjects are presented on different background scenes with changing positions on the screen. Hence, target groups and objects are presented concurrently (Correll et al., 2002). Alternatively, simplified versions of the FPST present pictures of faces paired with the target object superimposed or positioned side-by-side (Correll et al., 2014; Plant et al., 2005; Unkelbach et al., 2008). Regarding task instructions, participants are instructed to decide whether to "shoot" or "don't shoot" at a presented target subject. The IAT is a dual-categorization task presenting participants with two stimulus types: target groups (e.g., faces of Black and White male faces) and target attributes (here: objects representing threat and safety like guns versus tools) in a mixed, random order. Hence, target groups and objects are presented serially. Participants are instructed to categorize the displayed target stimuli, that is, target faces as Black or White and target objects as gun or tool. Crucially, in one IAT block one target group shares a response key with one target attribute

(i.e., Black/gun) and the other target group shares a response key with the other target attribute (i.e., White/tool). However, in another IAT block key pairings are switched (i.e., Black/tool, and White/gun). Hence, depending on target group and attribute pairing categorization is facilitated or hampered in performing the IAT. To recapitulate, these three measures inherently differ in their task instructions and stimuli presentation procedures. Regarding tasks instructions, participants are either instructed to identify the target object (WIT), to "shoot" or "not shoot" at a target subject (FPST), or to categorize target groups and attributes. Regarding stimulus presentations, stimuli are presented sequentially (WIT), concurrently (FPST), or serially (IAT).

These procedural differences may obscure correspondence among measures as different processes might contribute to responding. As mentioned before, MPT modeling allows to disentangle the joint contribution of multiple cognitive processes, as demonstrated with the PDP in Manuscripts 1 and 2. Nevertheless, other MPT models beside the PDP have been proposed to represent the cognitive process structure for implicit measures assessing racial bias. In total, we investigated seven MPT-models in Manuscript 3: the PDP (Payne, 2001) and its extended version including an additional guessing parameter (PDP+G; Bishara & Payne, 2009), the Quad model (Conrey et al., 2005) with its traditional specification assuming a different direction for stereotype-associations (i.e., Black males with guns, and White males with tools), the Quad model with an exploratory, egalitarian specification assuming the same direction for stereotype-associations (i.e., Black males with guns, and White males with guns), the Stroop model (Lindsay & Jacoby, 1994) and its extended version including an additional guessing parameter (Stroop+G; Bishara & Payne, 2009), and the Stereotype Misperception Task model (SMT; Krieglmeyer & Sherman, 2012). To investigate differences in cognitive process structure for these three measures, we estimated each model for each measure separately and assessed their model fit. In a subsequent step, we took the best fitting model for each measure and combined them in a single joint model. This joint model allows to investigate the correspondence between measures for parameters representing controlled and automatic processes as well as racial bias estimates calculated as the difference between automatic process parameters for Black and White male faces.

We conducted one experiment in Manuscript 3. In this experiment, participants completed a WIT, an FPST, and an IAT designed to assess threat-stereotypes towards Black and White males. The three measures were presented in a random order. To minimize procedural differences between measures, we aligned stimulus materials (i.e., the same Black and White male faces; the same drawings of guns and tools), number of critical trials ($N_{trials} = 240$), and response time limit (i.e., 700ms after target object onset) across measures. In addition, we added a neutral face prime consisting of the outline of a face to all measures. We did so to obtain more degrees of freedom in MPT-modeling. Importantly, this led to three face-pairings for the IAT (i.e., Black-White, Black-neutral, and White-neutral) which were administered in random order.

For all three measures racial bias emerged in error rates. However, correlation analysis for racial bias in error rates only revealed a weak correlation between the WIT and FPST (r = .18, p < .001), but no significant correlation emerged for the IAT with the other two measures. To investigate whether correspondence between measures was obscured by the influence of multiple cognitive processes, we applied MPT modeling to disentangle these processes.

First, we compared different MPT models for each measure to check for differences in cognitive process structure. Model estimation revealed acceptable fit across measures for control process-dominant MPT models (WIT: PDP and traditional Quad model; FPST: PDP; IAT: PDP, PDP+G and egalitarian Quad model). However, none of the automatic process-dominant MPT models (Stroop, Stroop+G, and SMT) revealed acceptable model fit. Model comparison indices revealed a preference for the PDP across all measures. Indicating that these measures can all be represented well by a relatively simple MPT model like the PDP. However, a more complex model like the Quad model that disentangles target identification and response conflict resolution can be viable for the WIT and IAT.

To investigate correspondence between measures, we submitted the PDP model for each measure in one joint model estimation. This allows to include estimates for parameter correlations directly in the modeling process (Heck, Arnold, & Arnold, 2018; Klauer, 2010). The joint model showed that the *C*-parameter, hence controlled processes leading to correct responding, correlated strongly among measures. Similarly, the *A*-parameter, hence the general tendency to respond with "gun" in comparison to "tool", correlated moderately to strongly among measures. Regarding racial bias estimates, calculated as the difference between *A*-parameters for Black and White male faces, they correlated moderately between the WIT and FPST (r = .38, 95%-BCI [.09 – .66]). But no correlation between the IAT with the other two measures emerged (both rs ≤ .16). Surprisingly, weak to moderate correlations emerged between the racial bias estimate of the IAT and *C*-parameters of the WIT (r = .26, 95%-BCI [.02 – .52]) and FPST (r = .34, 95%-BCI [.09 – .60]). Subsequent exploratory analysis revealed that this correlation was driven by responses to faces in the IAT-block directly contrasting Black and White male faces[4].

Overall, these results align well with previous findings reporting correspondence between racial bias estimates of the WIT and FPST (Ito et al., 2015) but low correspondence of the IAT with other racial bias measures (Ito et al., 2015; Olson & Fazio, 2003). This was even true when aligning measures across several procedural dimensions (i.e., stimulus materials, number of trials, and response time limits). The WIT and FPST are both configured to assess how race influences the fast behavioral response to guns and non-threatening target objects. Furthermore, both seminal articles introducing these paradigms explicitly refer to the behavior of police officers in shooting unarmed Black men (Correll et

[4] The exploratory joint PDP-model is available at: https://osf.io/2whze/

al., 2002; Payne, 2001; Payne & Correll, 2020). Although the WIT and FPST differ in stimulus presentation order (simultaneous vs. concurrent) and style (isolated stimulus in the center of the screen vs. with different backgrounds, full-body images, and different screen positions), they are able to assess a similar construct.

In contrast, the WIT and IAT differ on two crucial procedural dimensions: category salience and task instruction. In the standard WIT race is not made salient, but in the IAT participants are explicitly instructed to categorize faces by race in addition to categorizing target objects as guns or tools. Hence, race is made salient and directly contrasted in the IAT which may influence how stereotype associations are processed. For example, Olson and Fazio (2003) compared correspondence between an Evaluative Priming Task (EPT), a sequential priming paradigm which assesses positive evaluations toward Black and White males, with a Black-White evaluations IAT. The EPT and IAT only showed correspondence for participants who were manipulated to pay attention to race in the EPT. This indicates that race salience influences processing style in sequential priming paradigms, potentially shifting from an exemplar-based stereotype association to a race category-based stereotype association (Gawronski et al., 2010; Olson & Fazio, 2003). Hence, as race category is salient in the IAT but not in the WIT, this might explain the lack of correspondence in racial bias estimates between these measures.

Besides category salience, task instructions strongly differ between WIT and IAT. In the WIT participants are instructed to solely identify the target object, whereas in the IAT participants are performing dual-categorization of target objects and faces. This can lead participants to rely on recoding while responding. Recoding[5] means that stimulus categorization in the IAT does not rely on the four categories but simplifies them into two categories. For example, categorizing Black male faces as threatening together with gun targets, and White male faces as non-threatening together with tool targets reduces the IAT to a simple binary decision: threatening or not? This simplification might be easier for one IAT block (e.g., Black-guns) compared to the other block (e.g., Black-tools), resulting in less errors in one than the other block, which represents a racial bias effect (Meissner & Rothermund, 2013). In the PDP, recoding effects map on the A-parameter together with automatic stereotype-associations. This might explain the correspondence between the C-parameter of the WIT and FPST with the racial bias estimate for responses to faces in the Black-White IAT block, because recoding processes might be more pronounced for people with higher cognitive abilities (Meissner & Rothermund, 2013; von Stülpnagel & Steffens, 2010).

In conclusion, Manuscript 3 investigated the correspondence of three implicit measures configured to assess the threat-stereotypes for Black and White males: the WIT, FPST, and IAT. This comparison revealed that participants' capability to respond correctly generalizes across measures. In

[5] In accordance with Meissner and Rothermund (2013), I don't make any strong claim whether recoding is based on a deliberate strategy or on implicit learning.

addition, their tendency to respond with "gun" compared to "tool" persists across measures. However, measures diverged in their correspondence for the racial bias estimate. The WIT showed moderate correspondence to the FPST, both relying on race as a non-salient, goal-independent category that biases responses. In contrast, the WIT and IAT showed virtually no correspondence. This might be due to a difference in the assessed construct, as the WIT (and the FPST) probably measure a more spontaneous interference of stereotype associations based on the exemplar (Gawronski et al., 2010) and the IAT probably measures a more category-related stereotype construct due to category salience (De Houwer, 2001).

5. General Discussion

This dissertation examined the mechanisms and the nature of cognitive processes underlying racially biased weapon identification. Specifically, this line of research mainly focused on the Weapon Identification Task (WIT; Payne, 2001). The WIT is a sequential priming paradigm designed to assess the influence of racial threat-stereotypes on identifying threatening (e.g., guns) and non-threatening (e.g., tools) objects. For investigating underlying processes, we mainly implemented the Process Dissociation Procedure (PDP; Jacoby, 1991; Payne, 2001). The PDP is a dual-process model which allows to disentangle the influence of controlled and automatic processes in task performance.

In Manuscript 1 we applied a generalized version of the PDP to investigate whether racial bias in the WIT is mediated by target discrimination, response bias, or both. In two studies, we found that racial bias was mediated by response bias. However, racial bias was additionally mediated by target discrimination when participants were explicitly instructed to engage in racial profiling, hence using the primes' race as cue for weapon identification. Overall, these results suggest that face primes in the WIT create a second stream of information besides information extracted from the to-be-identified target. This additional information interferes with object identification, leading to a racially biased response when target discrimination is not successful. Nevertheless, when participants engage in racial profiling, target discrimination itself is biased towards "gun" responses given Black primes and towards "tool" for White primes. In sum, these findings suggest that both mechanisms can drive racially biased responding. Importantly, however, only response bias mediated racial bias in the standard version of the WIT.

In Manuscript 2 we investigated the nature and interplay of cognitive processes represented by four different psychological process models for the WIT (Klauer & Voss, 2008). In fact, these psychological process models were all instantiations of the PDP. Although these process models make the same predictions about the response patterns, they differ in their assumptions about the relative latencies of process branches. This allows us to compare models by applying response time-extended Multinomial Processing Tree modeling (MPT-RT; Heck & Erdfelder, 2016). Reanalysis of eight previously published data sets of the WIT showed the Default Interventionist Model (DIM) to perform best. The DIM postulates that a default response influenced by automatic stereotype associations is elicited at the onset of every trial. This default response is followed by controlled processes aiming to identify the displayed target object and to resolve possible conflicts if the default response and the to-be-identified target are incongruent. In sum, this means for the WIT that automatic stereotype associations play a key role in every trial of the WIT and that target identification, as well as possible response conflict resolution processes, are important for accurate responding.

In Manuscript 3, we investigated the correspondence of the WIT with other implicit measures (i.e., the First-Person Shooter Task, FPST, Correll et al., 2002; and the Implicit Association Test, IAT, Greenwald et al., 1998) configured to assess racial threat-stereotypes regarding Black and White males.

In addition, this study applied different types of Multinomial Processing Tree (MPT) models to look at the process-level correspondence of these measures. All three implicit measures were able to detect racial bias in error rates, but racial bias estimates based on error rates correlated only weakly between the WIT and FPST and not with the IAT. Regarding process level analysis, the three measures corresponded strongly among controlled processes which lead to correct responding, and moderately among automatic processes which represent the general response tendency towards "gun" compared to "tool". Regarding process-level based racial bias estimates, again they correlated moderately between the WIT and FPST but not with the IAT. However, the racial bias estimate of the IAT correlated with controlled process estimates of the WIT and FPST. In sum, this result pattern indicates that the WIT and FPST show moderate correspondence, as they both assess a spontaneous interference of exemplar-based racial threat-stereotype associations on object identification, whereas the WIT and IAT show low correspondence in racial bias estimates, as the IAT might assess more category-based racial treat-stereotype associations potentially due to category salience.

Taken together, the results of the three manuscripts broaden the understanding of how race shapes cognitive processes in differentiating a weapon from a non-threatening object. This was done by relying on model-based analysis for cognitive processes. Hence this thesis provides a deeper understanding for social cognition research, expanding knowledge from a social and a cognitive perspective as two sides of the same coin: on the one hand how social information shapes cognitive decision processes and on the other hand how cognitive processes influence social behavior.

5.1. Cognitive Processes and Factors Affecting Weapon Identification Bias

As outlined in the introduction section, weapon identification bias varies by different factors like, prime characteristics, experimental manipulations, and individual differences in participants. In the following I like to discuss, how these factors relate to the findings of this thesis[6].

Regarding prime characteristics, emotional expression of faces and social dimensions like age and gender influence the size of weapon identification bias besides race. This indicates that these features of prime characteristics are also processed early by participants and hence influence participants' response bias. The spontaneous and quick processing of emotions and facial characteristics when seeing a face is often observed (Hildebrandt et al., 2012; Kubota & Ito, 2007). However, the variability of bias in the WIT indicates that also stereotype associations for these features are concurrently elicited early on, which results in the variation of weapon identification bias by these features.

[6] The effects of task instruction and category salience will be discussed in the section on future directions.

Regarding experimental manipulations, a short response time window (Payne, 2001; Payne et al., 2002), cognitive depletion (Govorun & Payne, 2006), and social anxiety (Lambert et al., 2003), all diminish controlled processing in the WIT. Hence, these manipulations reduce the cognitive resources participants are able to dedicate to the task. As Manuscript 2 revealed, controlled processes in the WIT entail target discrimination and conflict resolution processes. However, the PDP, which was applied as an analytic framework in the studies above, does not allow to disentangle these two processes. Therefore, it is not clear whether these manipulations may attenuate target discrimination, conflict resolution, or both. For example, a short response time window may impede both processes, whereas cognitive depletion and social anxiety might only reduce conflict resolution but not target discrimination capabilities. However, alternative MPT models like the Quad model would allow to disentangle these controlled processes. Manuscript 3 demonstrated acceptable model fit for the Quad model applied to a WIT including additional neutral face primes. Overall, future research can rely on a similar setup of the WIT and test the differential influence of experimental manipulations on target discrimination and conflict resolution processes.

Similarly, individual differences in cognitive abilities and motivations are connected to correct responding in the WIT. So do cognitive executive function abilities (like inhibition) and internal motivations to control prejudice correlate positively with controlled responding in the WIT (Ito et al., 2015; Volpert-Esmond et al., 2020). Hence this aligns well with the DIM which features the important role of conflict resolution. Again, however, future research which likes to test the correspondence of individual capabilities and the ability to resolve conflict in the WIT may use different approaches. One approach could be using the Quad model as mentioned in the previous paragraph. Another approach could be using the MPT-RT model of the DIM from Manuscript 2 which allows one to estimate the difference in relative branch latencies. Specifically, it allows for the calculation of differences in response latencies between controlled processes branches with an incongruent default response to the target object versus controlled process branches with a congruent default response to the target object. In consequence, a small difference in relative latencies would indicate more effective conflict resolution capabilities. Overall, approaches like these would allow to investigate how participants' cognitive abilities and motivations correspond to participants' performance in general and to target discrimination and conflict resolution abilities in specific.

To conclude, factors affecting the WIT may relate to different cognitive processes in task performance. On the one hand, different prime characteristics relate to the automatic process. They are processed quickly and trigger early on stereotype associations, which interfere with participants' response efforts as an additional stream of information. On the other hand, experimental manipulations which hamper cognitive abilities as well as individual differences in cognitive abilities and motivations relate to the controlled process. Disentangling the role of target discrimination and conflict resolution processes

in future research would allow us to get a better picture of how they differentially influence task performance in the WIT. Overall, future research which takes into account more generalized versions of the PDP (Manuscript 1), response times extensions to the PDP (Manuscript 2), or alternative MPT models (Manuscript 3), would allow to us make more thorough, future investigations to clarify these assumed relationships.

5.2. Strengths and Limitations

This line of research demonstrates strong methodological procedures in experimental design and statistical analysis. Regarding experimental designs in Manuscript 1 and 3, we used previously validated stimulus materials which have already been used for the Weapon Identification Task (Phills et al., 2011; Rivers, 2017). The faces used were based on a high number of exemplars for each racial category and even provided the opportunity to include neutral face outlines as reference category. Target objects consisted of drawings of weapon and tool targets, which allowed us to control for perceptual features of the objects like color. Manuscript 2 relied on a reanalysis of eight published data sets based on different stimuli from different research teams. Consequently, this allowed us to compare whether the same process model emerges as the favored model from diverse data sets. Regarding statistical analysis, this line of research relied heavily on the Process Dissociation Procedure (PDP) as a model framework. This allowed on the one hand to rely on a well-established analytic framework for the WIT and related tasks that researchers are familiar with (e.g., Ito et al., 2015; Klauer et al., 2015; Payne, 2001). On the other hand, the PDP is a relatively simple model which can be expanded without drastic changes in its core structure (e.g., the generalized model in Manuscript 1; the MPT-RT extension in Manuscript 2). The PDP can be formalized in the framework of Multinomial Processing Tree (MPT) models. This enables conjoint parameter estimation and provides assessment of model fit. In addition, we used Bayesian hierarchical latent-trait MPT modeling (Heck, Arnold, & Arnold, 2018; Klauer, 2010), which allows to account for heterogeneity in samples and provides parameter estimates on the group and individual level. This, for example, allowed us to look at correlations between MPT parameters as done in Manuscript 3. Overall, this line of research relied on a methodologically sound approach, on which future studies can build on to investigate cognitive processes underlying weapon identification.

Nevertheless, this dissertation also comes with limitations. These were addressed in the respective manuscript. However, two limitations for Manuscript 1 and 3 will be discussed further in the following. In Manuscript 1 we applied a generalized version of the PDP. To achieve enough degrees of freedom for estimation we included an additional within-subject manipulation. This manipulation is believed to leave target discrimination unaffected while it likely effects overall response bias. This generalized version of the PDP was validated by Klauer et al. (2015) who relied on a pay-off manipulation to achieve more degrees of freedom. However, we relied on a base-rate manipulation for

the target objects to deliberately manipulate response bias but not target discrimination. Base-rate manipulations like these have been used before for similar tasks like the IAT (Conrey et al., 2005) and in recognition research (Buchner et al., 1995). However, in contrast to Klauer et al. (2015) our generalized PDP entailed zero degrees of freedom. In consequence, it could be expected to have no problems to fit the data and that model fit might be rather uninformative. However, that is not necessarily the case, because in a strict sense the PDP is not a fully saturated model. The PDP cannot account for data patterns where participants show more errors than correct responses for a specific target. This would correspond to a negative C-parameter estimate in the PDP. So, model fit is informative to rule out data patterns of below chance performance. Most importantly, although our generalized PDP has zero degrees of freedom, it is a feasible model to investigate the parameter estimates themselves and allows to test the parameter estimates via nested model comparisons including shrinkage parameters or equality restrictions[7].

In Manuscript 3, we compared three implicit measures configured to assess racial threat-stereotypes. For comparability reasons we aligned these measures on several dimensions. However, this led to slight variations of these measures mainly for the FPST and IAT which might attenuate the conclusions drawn from comparing the WIT to these two measures. Regarding the FPST, we presented target faces and objects at similar size and without surrounding context, though similar presentation styles have been implemented before for the FPST (see e.g., Plant et al., 2005; Unkelbach et al., 2008). Nevertheless, some researchers argue that such presentation style should rather be labeled as an FPS-type task (Payne & Correll, 2020). Regarding the IAT, two major changes were implemented in comparison to the standard paradigm. First, we implemented a response deadline which is not typically implemented for the IAT. Nevertheless, this has been done before and allowed to find racial bias effects (Calanchini et al., 2021; Conrey et al., 2005; Cunningham et al., 2001), which was also the case in our results. Second, we added a neutral face outline to provide sufficient degrees of freedom for MPT modeling. This resulted for the IAT in an additional set of blocks to accommodate this third category (i.e., Black-White, Black-neutral, White-neutral). However, this expanded IAT format is similar to an existing version of the IAT: the multi-category IAT (Axt et al., 2014). In addition, exploratory analysis revealed that the reliable racial bias effect emerged in the Black-White IAT-block resulting in the same conclusions as integrating data across all IAT-blocks.

[7] As a side note, a similar model comparison strategy based on the PDP was applied in the first model comparison of Manuscript 2. All the models had zero degrees of freedom but differed in their assumptions on the order of relative branch latencies which were implemented via shrinkage parameters, restricting the plausible data space for each model. These restrictions helped to distinguish between these models, ruling out models which were not able to represent fast automatic and slow controlled branch latencies.

To sum up, these limitations for the generalized PDP model specification (Manuscript 1) and for task adaptations (Manuscript 1 and 3) should be kept in mind for interpreting our findings. Nevertheless, experimental designs and assumptions were based on previous, methodological sound research using similar modifications. Going on, our findings provide a strong basis for future research investigating the role of cognitive processes in weapon identification.

5.3. Future Directions

5.3.1. Alternative Modeling Approaches

Throughout the three manuscripts we relied on MPT modeling in general and on the PDP in specific to investigate underlying cognitive processes. This is in line with previous research, which typically applies the PDP to disentangle the contribution of controlled and automatic processes. However, other formal modeling approaches for cognitive processes can be applied to the WIT.

Regarding the MPT-framework, the Quad model, which separates the influence of target discrimination and conflict resolution, has been used previously (Conrey et al., 2005) and provided also acceptable fit in Manuscript 3. Nevertheless, the Quad model comes with the downside that due to its higher number of parameters, it cannot be applied easily to the standard version of the WIT. As the basic paradigm with two types of faces (i.e., Black and White) and two types of objects (i.e., guns and tools) only provide four degrees of freedom. In consequence, to apply the Quad model meaningfully to the WIT additional experimental conditions providing more degrees of freedom compared to the standard procedure of the WIT are necessary (e.g., including neutral face primes).

Conventional MPTs rely on discrete data like response frequencies for modeling. However, more recent approaches allow to integrate response latencies, for example response time-extended MPT modeling (MPT-RT; Heck & Erdfelder, 2016) which we used in Manuscript 2 to estimate relative response latencies of branches. Yet, other approaches integrating response times in MPT models exist, like RT-MPT models (Klauer & Kellen, 2018) or Generalized Processing Tree models (GPT; Heck, Erdfelder, & Kieslich, 2018). Additional to the basic categorical structure of MPTs, RT-MPTs allow to estimate the response time distribution of each process separately, whereas GPTs allow to estimate the distribution of a continuous variable (like e.g., response times or neurophysiological variables) belonging to an entire branch. Hence, they would allow to obtain a more fine-grained view on the role of response latencies. In addition, the GPT would come with the possibility to integrate neurophysiological data. This would allow for example to investigate whether for the Default Interventionist Model of Manuscript 2 the activity of specific brain regions (Cunningham et al., 2004; Rivera-Rodriguez et al., 2021) maps on the expected branches. Therefore, this would enable to investigate whether the activity of brain regions associated with response conflict resolution are reflected in branches which include incongruent information proposed by the automatic default response and by the target to-be-identified. However,

these models come with the downside of a priori assumptions about the distribution of the continuous variables and estimation itself is in general less stable and more complex.

Besides MPT models, other formal modeling approaches like Diffusion Models (DM) have been discussed and applied to the WIT (Klauer & Voss, 2008; Todd et al., 2021). A DM is an evidence-accumulation model integrating the joint contributions of response frequencies and response times (Ratcliff, 1978; Ratcliff et al., 2016; Voss & Voss, 2007). A DM decomposes four determinants in task performance: the initial response bias towards guns or tools, the quality of information extracted from the target stimulus (so-called drift rate), the non-decision time (like motor response time and encoding), and the amount of evidence required to make a decision (so-called threshold separation). In their analysis, Todd et al. (2021) found that racial bias was reflected in initial response bias but not in the drift rate, analogous to our finding in Manuscript 1. Hence, although using a different formal modeling approach, they arrived at similar conclusions.

In sum, future research on the WIT can benefit from different modeling approaches to get a better, multi-perspective insight on cognitive processes underlying task performance, while combining their strengths, and eventually achieving converging evidence from different approaches (Klauer & Voss, 2008). Furthermore, increasing computational power and scientific open access practices for modeling packages (Hartmann et al., 2020; Heck, Arnold, & Arnold, 2018; Heck, Erdfelder, & Kieslich, 2018) and analytic code (Todd et al., 2021) facilitates the usage of formal modeling in data analysis. In addition, statistical solutions for hierarchical model implementations allow to integrate participants' heterogeneity in modeling and to investigate individual differences (Heck, Arnold, & Arnold, 2018; Klauer, 2010).

5.3.2. First-Person Shooter Task

This dissertation mainly relies on the WIT as a task to investigate how racial threat-stereotypes influence cognitive processes in identifying weapons. However, the First-Person Shooter Task (FPST; Correll et al., 2002) was also developed to measure the influence of racial threat-stereotypes on behavior around the same time as the WIT. Both tasks are seen as valid instruments to assess this racial bias (Payne & Correll, 2020). As we have shown in Manuscript 3, these two measures revealed a moderate correspondence in their racial bias estimates.

However, it remains indetermined whether the nature and interplay of cognitive processes, which lead to racially biased behavior, are the same across measures because of their significant procedural differences. The FPST presents the race stimulus concurrent with the target object, embeds stimuli in varying background scenes, and instructs participants to decide whether to "shoot" or "don't shoot" at a presented person. This challenges the generalizability of the findings of Manuscript 1 and 2 to the FPST. For example, regarding Manuscript 2, is the preference of the DIM for the WIT transferable to the FPST? Because the concurrent presentation of race stimulus and target object in

the FPST might not necessarily mean that race biases an initial default response. On the contrary, the search character of the task might lead to trials for which participants do not even process race information. Likewise, are the findings of Manuscript 1 transferable to the FPST? Hence, is racial bias in the standard paradigm mediated via response bias and not discrimination bias? In that regard, diffusion modeling for the FPST revealed that in standard versions of the FPST, race moderated drift rates for armed and non-armed persons (mirroring target discrimination) but did not moderate initial response bias (Correll et al., 2015; Pleskac et al., 2018). This influence of race on information extraction from the target would suggest that controlled processes are influenced by prime and target condition in the standard version of the FPST, but not automatic response bias. Hence, this reverses the findings compared to the standard version of the WIT in Manuscript 1. To substantiate this point, as mentioned earlier, Todd et al.'s (2021) diffusion modeling analysis for the WIT resulted also in an effect of race on the initial response bias but not on the drift rate. In conclusion, although both measures reliably assess racial threat-stereotypes, they seem to be based on different mechanisms driving racially biased responding.

5.3.3. The Influence of Race Category Salience

As seen in Manuscript 1 and as discussed as potential lack of correspondence between the WIT and IAT in Manuscript 3, race category salience may influence mechanisms driving racial bias in the WIT and the correspondence between implicit measures. Race category salience typically strengthens racial bias. For example, instructions to rely on or to ignore prime race information enhance racial bias in the WIT in comparison to the standard instruction to disregard the prime without mentioning race (Payne et al., 2002; Manuscript 1). Similarly, category salience manipulations like categorizing faces beforehand or counting faces by category during the task reveal an enhanced racial bias effect when attending to race but an attenuated effect when attending to another social category like age (Jones & Fazio, 2010; Todd et al., 2021) or a non-substantive category like the color of a dot superimposed on a face (Ito & Tomelleri, 2017; Todd et al., 2021). However, the question is, how does race category salience modulate the different cognitive processes in weapon identification. In Manuscript 1 racial profiling led to biased target discrimination and a descriptively more pronounced effect on response bias[8]. Todd et al.'s (2021) category salience manipulation induced by a previous face sorting task (i.e., by race, age, or

[8] Payne et al., 2002 report a more pronounced effect of race on automatic processes (response bias) in the racial profiling and the ignore race instruction condition but no effect of race on controlled processes (target discrimination). However, in the conventional version of the PDP a prime-target interaction of the C-parameter is not accounted for and to that effect the influence of a prime-target interaction of the C-parameter erroneously maps on the A-parameters as discussed by Klauer et al. (2015). Hence, out of Payne et al.'s (2002) findings one cannot conclude that the racial profiling or the ignore race instructions do not influence controlled processes.

superimposed dot color) revealed in their diffusion model analysis an enhanced racial bias effect on initial response bias when race category was salient in comparison to when non-racial categories were salient. Additionally, it revealed a main effect of race on the drift rate in the race category salience condition (i.e., faster evidence accumulation after Black faces for gun and tool targets) but no effect in the non-racial category salience conditions. Hence, race category salience seems to boost response bias and to moderate the effect of race on controlled processes in the WIT.

In contrast, a study by Stewart and Payne (2008) showed that if participants use implementation intentions (i.e., predefined if-then action plans) to think "safe" when they see a Black face, weapon identification bias is attenuated. Implementation intentions like these are twofold, on the one hand, they make race category salient to participants, but on the other hand, they teach participants a practical strategy to circumvent racial bias. Stewart and Payne (2008) report that implementation intentions attenuate racial bias in automatic process parameter estimates of the PDP but no difference in controlled process parameter estimates. However, they applied the conventional version of the PDP (i.e., A- and C-parameters vary by race but are restricted across target objects), which cannot distinguish between response bias or discrimination bias as mechanisms for racial bias. In consequence, it remains undetermined whether implementation intentions help to attenuate racial bias mediated via response bias, discrimination bias, or both. Or even intensify the influence of race on one mechanism but dimmish it on the other. Overall, future research on the influence of race category salience on WIT performance should consider that controlled processes might vary by prime race and target object. Therefore, study designs should include analytic models which are able to account for racial bias in automatic and controlled processes.

Looking at research on Evaluative Priming Tasks (EPT), which are sequential priming tasks assessing the positive and negative evaluations of Black and White male faces (Fazio et al., 1995), they find that category salience moderates the effect of racial evaluations. Furthermore, the effect of category salience may rely on specific mechanisms rooted in task procedure and might be reflected by controlled processes. Gawronski et al. (2010) compared the influence of category salience on racial bias in the EPT and the Affect Misattribution Procedure (AMP; Payne, Cheng et al.,2005) for faces varying by race (i.e., Black vs. White) and age (i.e., young vs. old). Category salience modulated bias in the EPT, when race was made salient racial bias was elicited and when age was made salient age bias was elicited. However, for the AMP race and age bias were present independent of respective category salience. Gawronski et al. (2010) argued that the AMP relies on diffuse affective states leading to misattribute evaluative judgments, whereas the EPT captures information that is in the attentional focus of participants which then interferes in task responding. Hence, category salience should increase attentional focus which is related to controlled processing. Two mechanisms, as already discussed in Manuscript 1, might be plausible how race information might interfere with controlled processes: the

first one referring to discrimination bias due to racially biased perception and information extraction from the target object (Klauer et al., 2015), or the second one referring to response conflicts between a default response tendency congruent or incongruent with the to-be-identified target, hence requiring more or less conflict resolution capabilities (Gawronski et al., 2010). However, the method used in Manuscript 1 is not able to test between these two plausible mechanisms and further research is needed to investigate which of these mechanisms, or perhaps both, are driving task performance when the critical category is salient.

In a similar vein, as discussed in Manuscript 3, race category salience might explain the lack of correspondence between the WIT as sequential priming task and the IAT as dual-categorization task. Due to the categorization instruction character of the IAT, stereotype associations are seen as category-based (De Houwer, 2001), whereas in sequential priming tasks associations are seen as exemplar-based (Livingston & Brewer, 2002). In accordance, Olson and Fazio (2003) found that a Black-White EPT and IAT show correspondence when race category was made salient, emphasizing category-based associations in the EPT (Gawronski et al., 2010), but no correspondence when not made salient, due to exemplar-based associations for the EPT. Hence, future research should systematically investigate the influence of race category salience on the process-level correspondence of the WIT to other implicit measures.

6. Conclusion

The aim of this dissertation is to get a deeper understanding of cognitive processes leading to racially biased misidentification of non-threatening objects as weapons. Overall, this line of research investigated racially biased behavior on a cognitive process-level. It showed that race spontaneously triggers threat-stereotypes interfering with weapon identification, which eventually need to be overcome for correct object identification. Furthermore, racial profiling and race category salience can additionally bias participants' higher order processes involved in target discrimination. Overall, this thesis contributes to a deeper understanding of how racially biased behavior is produced by different processes. Furthermore, the research included in the thesis helps to find ways to counteract racially biased behavior and to improve overall decision making in identifying threatening and non-threatening objects.

7. References

Amon, M. J., & Holden, J. G. (2016). *Fractal scaling and implicit bias: A conceptual replication of Correll (2008).* In A. Papafragou, D., Grodner, D., Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1553–1558). Austin, TX: Cognitive Science Society. https://cognitivesciencesociety.org/wp-content/uploads/2019/03/cogsci2016_proceedings.pdf

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, *25*(9), 1804–1815. https://doi.org/10.1177/0956797614543801

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*, 668–688. https://doi.org/10.3758/s13428-013-0410-6

Bishara, A. J., & Payne, B. K. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology*, *45*(3), 524–534. https://doi.org/10.1016/j.jesp.2008.11.002

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(5), e1501. https://doi.org/10.1002/wcs.1501

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General*, *124*(2), 137–160. https://doi.org/10.1037/0096-3445.124.2.137

Calanchini, J., Meissner, F., & Klauer, K. C. (2021). The role of recoding in implicit social cognition: Investigating the scope and interpretation of the ReAL model for the implicit association test. *PloS one*, *16*(4), e0250068. https://doi.org/10.1371/journal.pone.0250068

Calanchini, J., Sherman, J. W., Klauer, K. C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of IAT performance. *Personality and Social Psychology Bulletin*, *40*(10), 1285–1296. https://doi.org/10.1177/0146167214540723

Conrey, 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*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, *24*(3), 212–232. https://doi.org/10.1177/1088868320911325

Correll, J. (2008). 1/f noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology*, *94*(1), 48–59. https://doi.org/10.1037/0022-3514.94.1.48

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass*, *8*(5), 201–213. https://doi.org/10.1111/spc3.12099

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: how stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology*, *108*(2), 219–233. https://doi.org/10.1037/pspa0000015

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, *15*(12), 806–813. https://doi.org/10.1111/j.0956-7976.2004.00760.x

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*(2), 163–170. https://doi.org/10.1111/1467-9280.00328

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*(6), 443–451. https://doi.org/10.1006/jesp.2000.1464

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1–20. https://doi.org/10.3758/LB.37.1.1

Evans, J. S. B. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339. https://doi.org/10.1080/13546780601008825

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145030

Fish, J., & Syed, M. (2020). Racism, discrimination, and prejudice. *The Encyclopedia of Child and Adolescent Development*, 1–12. https://doi.org/10.1002/9781119171492.wecad464

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, *14*(4), 574–595. https://doi.org/10.1177/1745691619826015

Gawronski, B., & Brannon, S. M. (2019). What is cognitive consistency, and why does it matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a Pivotal Theory in Psychology* (pp. 91–116). American Psychological Association. https://doi.org/10.1037/0000135-005

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects? *Cognition and Emotion*, *24*(6), 1008–1025. https://doi.org/10.1080/02699930903112712

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition*, *38*(Supplement), s1–s25. https://doi.org/10.1521/soco.2020.38.supp.s1

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, *44*(1), 164–172. https://doi.org/10.1016/j.jesp.2007.01.002

Govorun, O., & Payne, B. K. (2006). Ego-depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, *24*(2), 111–136. https://doi.org/10.1521/soco.2006.24.2.111

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hartmann, R., Johannsen, L., & Klauer, K. C. (2020). rtmpt: An R package for fitting response-time extended multinomial processing tree models. *Behavior Research Methods*, *52*, 1313–1338. https://doi.org/10.3758/s13428-019-01318-x

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*, 264–284. https://doi.org/10.3758/s13428-017-0869-7

Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. https://doi.org/10.3758/s13423-016-1025-6

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika*, *83*, 893–918. https://dx.doi.org/10.1007/s11336-018-9622-0

Hildebrandt, A., Schacht, A., Sommer, W., & Wilhelm, O. (2012). Measuring the speed of recognising facially expressed emotions. *Cognition & Emotion*, *26*(4), 650 – 666. https://doi.org/10.1080/02699931.2011.602046

Huntsinger, J. R., Sinclair, S., & Clore, G. L. (2009). Affective regulation of implicitly measured stereotypes and attitudes: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *45*(3), 560–566. https://doi.org/10.1016/j.jesp.2009.01.007

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial

bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Ito, T. A., & Tomelleri, S. (2017). Seeing is not stereotyping: The functional independence of categorization and stereotype activation. *Social Cognitive and Affective Neuroscience*, *12*(5), 758–764. https://doi.org/10.1093/scan/nsx009

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jones, C. R., & Fazio, R. H. (2010). Person categorization and automatic racial stereotyping effects on weapon identification. *Personality and Social Psychology Bulletin*, *36*(8), 1073–1085. https://doi.org/10.1177/0146167210375817

Kahn, K. B., & Martin, K. D. (2020). The social psychology of racially biased policing: Evidence-based policy responses. *Policy Insights from the Behavioral and Brain Sciences*, *7*(2), 107–114. https://doi.org/10.1177/2372732220943639

Kidder, C. K., White, K. R., Hinojos, M. R., Sandoval, M., & Crites Jr, S. L. (2018). Sequential stereotype priming: A meta-analysis. *Personality and Social Psychology Review*, *22*(3), 199–227. https://doi.org/10.1177/1088868317723532

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221. https://doi.org/10.1037/xge0000044

Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology*, *82*, 111–130. https://doi.org/10.1016/j.jmp.2017.12.003

Klauer, K. C., & Voss, A. (2008). Effects of race on responses and response latencies in the weapon identification task: A test of six models. *Personality and Social Psychology Bulletin*, *34*(8), 1124–1140. https://doi.org/10.1177/0146167208318603

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*(4), 215–229. https://doi.org/10.1016/j.jmp.2004.03.002

Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, *103*(2), 205–224. https://doi.org/10.1037/a0028764

Kubota, J. T., & Ito, T. A. (2007). Multiple cues in social perception: The time course of processing race and facial expression. *Journal of Experimental Social Psychology, 43(5), 738–752.* https://doi.org/10.1016/j.jesp.2006.10.023

Kubota, J. T., & Ito, T. A. (2014). The role of expression and race in weapons identification. *Emotion*, *14*(6), 1115–1124. https://doi.org/10.1037/a0028764

Lambert, A. J., Payne, B. K., Jacoby, L. L., Shaffer, L. M., Chasteen, A. L., & Khan, S. R. (2003). Stereotypes as dominant responses: on the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology*, *84*(2), 277–295. https://doi.org/10.1037/0022-3514.84.2.277

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: the relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 219–234. https://doi.org/10.1037/0096-1523.20.2.219

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*(1), 5–18. https://doi.org/10.1037/0022-3514.82.1.5

Lundberg, G. J., Neel, R., Lassetter, B., & Todd, A. R. (2018). Racial bias in implicit danger associations generalizes to older male targets. *PloS one*, *13*(6), e0197398. https://doi.org/10.1371/journal.pone.0197398

Madurski, C., & LeBel, E. P. (2015). Making sense of the noise: Replication difficulties of Correll's (2008) modulation of 1/f noise in a racial bias task. *Psychonomic Bulletin & Review*, *22*, 1135–1141. https://doi.org/10.3758/s13423-014-0757-4

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *50*(3), 629–658. https://doi.org/10.1111/nous.12089

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*(1), 45–69. https://doi.org/10.1037/a0030734

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42–54. https://doi.org/10.3758/BRM.42.1.42

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science*, *14*(6), 636–639. https://doi.org/10.1046/j.0956-7976.2003.psci_1477.x

Payne, B. K. (2001). Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192. https://doi.org/10.1037/0022-3514.81.2.181

Payne, B. K. (2005). Conceptualizing control in social cognition: how executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology*, *89*(4), 488–503. https://doi.org/10.1037/0022-3514.89.4.488

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*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology*, *62*, 1–50. https://doi.org/10.1016/bs.aesp.2020.04.001

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology*, *38*(4), 384–396. https://doi.org/10.1016/S0022-1031(02)00006-9

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology*, *41*(1), 36–47. https://doi.org/10.1016/j.jesp.2004.05.001

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248. https://doi.org/10.1016/j.jesp.2004.05.001

Phills, C. E., Kawakami, K., Tabi, E., Nadolny, D., & Inzlicht, M. (2011). Mind the gap: Increasing associations between the self and blacks with approach behaviors. *Journal of Personality and Social Psychology*, *100*(2), 197–210. https://doi.org/10.1037/a0022159

Plant, E. A., Peruche, B. M., & Butz, D. A. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology, 41*(2), 141–156. https://doi.org/10.1016/j.jesp.2004.07.004

Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review*, *25*, 1301–1330. https://doi.org/10.3758/s13423-017-1369-6

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*(3), 318–339. https://doi.org/10.1037/0033-295X.95.3.318

Rivera-Rodriguez, A., Sherwood, M., Fitzroy, A. B., Sanders, L. D., & Dasgupta, N. (2021). Anger, race, and the neurocognition of threat: attention, inhibition, and error processing during a weapon identification task. *Cognitive Research: Principles and Implications*, *6*, 1–27. https://doi.org/10.1186/s41235-021-00342-w

Rivers, A. M. (2017). The weapons identification task: Recommendations for adequately powered research. *Plos one*, *12*(6), e0177857. https://doi.org/10.1371/journal.pone.0177857

Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial

processing tree models: A tutorial. *Psychological Methods*. Advance Online Publication. https://doi.org/10.1037/met0000561

Sherman, J. W. (2006). On building a better process model: It's not only how many, but which ones and by which means? *Psychological Inquiry*, *17*(3), 173–184. https://doi.org/10.1207/s15327965pli1703_3

Sherman, J. W., Klauer, K. C., & Allen, T. J. (2010). Mathematical modeling of implicit social cognition: The machine in the ghost. In B. Gawronski & B. K. Payne (Eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications* (pp. 156–175). New York, NY: Guilford Press. https://escholarship.org/uc/item/17b933nt

Subra, B., Muller, D., Fourgassie, L., Chauvin, A., & Alexopoulos, T. (2018). Of guns and snakes: testing a modern threat superiority effect. *Cognition and Emotion*, *32*(1), 81–91. https://doi.org/10.1080/02699931.2017.1284044

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*(10), 1332–1345. https://doi.org/10.1177/0146167208321269

Thiem, K. C., Neel, R., Simpson, A. J., & Todd, A. R. (2019). Are Black women and girls associated with danger? Implicit racial bias at the intersection of target age and gender. *Personality and Social Psychology Bulletin*, *45*(10), 1427–1439. https://doi.org/10.1177/0146167219829182

Todd, A. R., Johnson, D. J., Lassetter, B., Neel, R., Simpson, A. J., & Cesario, J. (2021). Category salience and racial bias in weapon identification: A diffusion modeling approach. *Journal of Personality and Social Psychology*, *120*(3), 672–639. https://doi.org/10.1037/pspi0000279

Todd, A. R., Thiem, K. C., & Neel, R. (2016). Does seeing faces of young black boys facilitate the identification of threatening stimuli? *Psychological Science*, *27*(3), 384–393. https://doi.org/10.1177/0956797615624492

Unkelbach, C., Forgas, J. P., & Denson, T. F. (2008). The turban effect: The influence of Muslim headgear and induced affect on aggressive responses in the shooter bias paradigm. *Journal of Experimental Social Psychology*, *44*(5), 1409–1413. https://doi.org/10.1016/j.jesp.2008.04.003

Volpert-Esmond, H. I., Scherer, L. D., & Bartholow, B. D. (2020). Dissociating automatic associations: Comparing two implicit measurements of race bias. *European Journal of Social Psychology*, *50*(4), 876–888. https://doi.org/10.1002/ejsp.2655

von Stülpnagel, R., & Steffens, M. C. (2010). Prejudiced or just smart? Intelligence as a confounding factor in the IAT effect. *Zeitschrift für Psychologie/Journal of Psychology*, *218*(1), 51–53. https://doi.org/10.1027/0044-3409/a000008

A Copies of Manuscripts

The Nature of Racial Bias in the Weapon Identification Task: Discrimination Bias, Response Bias, or Both?

Ruben Laukenmann[1] and Edgar Erdfelder[1]

[1]University of Mannheim

Author Note

Ruben Laukenmann https://orcid.org/0000-0002-4780-4845

Edgar Erdfelder https://orcid.org/0000-0003-1032-3981

Manuscript preparation and data collection was supported by the Research Training Group "Statistical Modeling in Psychology" (SMiP), funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, GRK 2277), and by a Doctoral Research Fellowship from the Stiftung der Deutschen Wirtschaft gGmbH (sdw, Foundation of German Business), funded by the German Federal Ministry of Education and Research (awarded to RL).

The authors are very grateful to C. E. Phills and A. M. Rivers for sharing the stimulus materials we used for our studies.

Correspondence concerning this article should be addressed to Ruben Laukenmann or Edgar Erdfelder, Senior professorship for Cognitive Psychology, A5, B 118, University of Mannheim, D-68159 Mannheim, Germany (email: ruben.laukenmann@psychologie.uni-mannheim.de or erdfelder@uni-mannheim.de).

Online supplementary materials, including the pre-registration of both studies reported, are available at: https://osf.io/4yxv9/?view_only=aa0ecf147152463ea657b24375f6c8f2

We have no conflicts of interest to disclose.

Abstract

In the Weapon Identification Task (WIT), the effect of racial primes (i.e., White vs. Black male faces) on visual discrimination between weapons and innocuous objects (i.e., guns vs. tools) is assessed. The typical finding is that participants more often misclassify innocuous objects as weapons after seeing a Black compared to a White male face. Discrimination bias and response bias have been proposed as explanations of this effect. Discrimination bias may result from effects of the preceding prime's race on information extraction and perceptual construal of the target object. Response bias, in contrast, may be induced by prime-triggered threat stereotypes that create an additional stream of information depending on race. In two experiments, we manipulated base-rates of target objects in the WIT and used a generalized process dissociation model to disentangle discrimination and response biases as determinants of racial bias in the WIT. Our results suggest that participants typically discriminate objects uninfluenced by primes whereas response bias systematically distorts WIT judgments. This notwithstanding, object discrimination is affected by race when participants are explicitly instructed to use racial information for responding. Hence, if participants choose to attend to the information in the preceding prime, this may deteriorate the identification of the target in the WIT. Our results suggest two strategies to counteract racial biases in spontaneous behavior: (1) Reducing response bias by enhancing the capacity to inhibit automatic threat-based associations and (2) facilitating unbiased target discrimination by reducing race salience.

Abstract word count: 238

Keywords: weapon identification task, discrimination bias, response bias, process dissociation procedure, multinomial processing tree modeling, racial bias

Word count: 8045 (excluding references, figures, tables, and footnotes)

The Nature of Racial Bias in the Weapon Identification Task: Discrimination Bias, Response Bias, or Both?

If an object is presented along with a person, the person's race can influence how good people are at identifying this object (Payne, 2001; Payne & Correll, 2020; Rivers, 2017). One of the most prominent procedures to investigate the influence of race on identifying potentially dangerous objects is the Weapon Identification Task (WIT; Payne, 2001). The WIT is a sequential priming procedure in which participants are instructed to identify an object as either a weapon (i.e., a gun) or an innocuous object (i.e., a tool). These objects are preceded by either a Black or White male face prime, typically inducing a racially biased response pattern. Participants tend to misidentify innocuous objects more often as weapons if they are preceded by a Black face prime. Furthermore, participants tend to identify weapon targets faster if they are preceded by a Black face prime. This result pattern is known as weapon identification bias (Payne, 2001; Rivers, 2017; Todd et al., 2021).

Mechanisms of Racially Biased Responding

Although weapon identification bias has been observed reliably, the mechanism leading to this bias has been a matter of debate (Klauer & Voss, 2008; Klauer et al., 2015; Payne et al., 2005; Todd et al., 2021). Two mechanisms have been proposed to drive racial bias. One possible mechanism is discrimination bias resulting from prime-influenced information extraction during perceptual construal of the target object. A second candidate is response bias induced by prime-triggered threat stereotypes that create an additional stream of information which varies between primes (Klauer et al., 2015; Klauer & Voss, 2008; Payne et al., 2005). The aim of our research is to assess both mechanisms empirically.

If discrimination drives racial bias in the WIT, then prime race should affect the perception and interpretation of parts of the target objects. Specifically, when participants try hastily to resolve perceptual ambiguity during object discrimination, the prime race provides context cues that might foster misperception of the target object. For instance, a metal tube of a target object might be interpreted as the barrel of a gun following a Black prime and as the shaft of a screwdriver following a White prime (Klauer et al., 2015; Klauer & Voss, 2008; Payne et al., 2005). This effect should not

necessarily persist when participants are given more time for responding. Supporting this account, Payne et al. (2005) asked participants for a second object classification judgment after providing their original response. Notably, in trials with incorrect initial object classifications, they were able to report the correct object when given sufficient time afterwards. Payne and collaborators concluded that target discrimination is not influenced by race, at least if participants are given sufficient time for responding. Additional model-based studies reported mixed results. While some authors found evidence for racially biased object perceptions (Klauer et al., 2015), others did not (Todd et al., 2021).

If racial bias is mediated by response bias, racial stereotypes elicited by the prime create a second stream of information besides the actual perceptual information. This results in different response biases triggered by racial stereotypes, that is, a stronger preference to respond "gun" after seeing a Black face and to respond "tool" following a White face. This aligns well with recent research suggesting that prime race influences response bias from the outset of each trial (Klauer & Voss, 2008; Laukenmann et al., 2023; Todd et al., 2021). Importantly, because response bias is based on a secondary stream of information, it only determines the response if object discrimination fails within the time window provided for responding.

Both accounts propose plausible mechanisms that might lead to racial bias in the WIT and are not mutually exclusive. Most important for our present study, both accounts map on different parameters of the Process Dissociation Procedure (PDP), a WIT measurement model frequently used to disentangle controlled and automatic influences on responding. In the following section, we introduce the PDP and show how possible influences of race primes on object discrimination and response bias can be disentangled in this framework.

Process Dissociation Procedure and Multinomial Processing Tree Modeling

The PDP is a widely used dual-process model to disentangle the contributions of automatic and controlled processes in the WIT (Bishara & Payne, 2009; Huntsinger et al., 2009; Ito et al., 2015; Klauer et al., 2015; Payne, 2001; Payne, 2005). The controlled process parameter reflects a participant's latent ability to identify the target correctly, whereas the automatic process parameter

reflects a participant's latent preference to respond with gun rather than tool (Klauer & Voss, 2008; Laukenmann et al., 2023; Payne, 2001).

Figure 1 illustrates the PDP in the form of two processing tree diagrams, one for each target condition (i.e., gun and tool). The left side indicates the to-be-identified target object, followed by the possible processing branches (and their associated probabilities) that lead to certain responses in the WIT, as indicated on the right side of the trees. For the gun target tree, for example, the controlled process succeeds with probability C, resulting in the correct response *gun*. When the controlled process fails with probability $(1 - C)$, the automatic process triggers the response *gun* with probability A and the response *tool* with the complementary probability $(1 - A)$. For the tool target tree, the processing branches are analogous, except that the controlled process results in the correct response *tool*.

Insert Figure 1 about here

Parameters of process dissociation models such as those illustrated in Figure 1 can be estimated using Multinomial Processing Tree (MPT) modeling (for a tutorial, see Schmidt, Erdfelder, & Heck, 2023). MPT models have previously been used to estimate PDP parameters for the WIT (Bishara & Payne, 2009; Klauer et al. 2015, Laukenmann et al., 2023). Compared to alternative estimation methods based on simple algebraic equations (Payne, 2001), MPT modeling enables conjoint estimation of all relevant process parameters C and A including standard errors, along with the assessment of model fit and calculation of information criteria for purposes of selecting among several candidate models.

Hypothesis Testing and Auxiliary Assumptions

The two possible mechanisms hypothesized to underlie racial bias in the WIT – discrimination bias and response bias – are associated with different parameters of the PDP. Specifically, while response bias maps on the automatic process, discrimination bias maps on the controlled process. Hence, the PDP can be used to assess both candidate mechanisms (Klauer et al., 2015).

More precisely, if stereotype-influenced response bias is at work, this should cause a higher rate of gun responses following Black face primes, given failure of the controlled process. This will be mirrored in a larger automatic process parameter A for the Black (B) than for the White face prime (W) condition (i.e., $A_B > A_W$). According to the response-bias hypothesis, the difference $A_B - A_W$ measures racial bias in the WIT (Klauer et al., 2015; Ito et al., 2015; Payne, 2001). As an alternative racial bias measure, a shrinkage parameter s_A that represents the ratio A_W/A_B, can be used to implement the order restriction $A_W \leq A_B$ (Knapp & Batchelder, 2004), such that $A_W = s_A \cdot A_B$ with $0 \leq s_A \leq 1$.

In contrast, if discrimination bias is at work, the prime race affects object discrimination. Hence, the controlled process parameter C is expected to vary as a function of both prime and target conditions. In the Black prime condition, the prime is predicted to enhance target discrimination for guns (G) compared to tools (T), resulting in a larger controlled process parameter for gun than for tool targets (i.e., $C_{BG} > C_{BT}$). In the White prime condition, in contrast, the prime is predicted to have opposite effects, resulting in $C_{WG} < C_{WT}$ (Klauer & Voss, 2008; Klauer et al., 2015). Again, these predictions can be formalized in terms of parameter differences, ratios, or shrinkage parameters. Note that the standard version of the PDP cannot accommodate this prime by target interaction because the parameters C_B and C_W are equated across target conditions (i.e., $C_{BG} = C_{BT} = C_B$ and $C_{WG} = C_{WT} = C_W$ is assumed in the standard PDP). Hence, to evaluate whether response bias, discrimination bias, or both influence racially biased responding in the WIT, a reparameterization of the PDP model is necessary that allows the C-parameter to vary between both prime and target conditions. Note, however, that developing such a generalized PDP variant from the model illustrated in Figure 1 would result in an overparameterized model: It would include six parameters (i.e., four C-parameters and two A-parameters) but provide only four independent category probabilities (i.e., probabilities of correct responses for each of the four prime-target combinations under investigation). Hence, such a model would be technically nonidentifiable so that parameters cannot be estimated.

More degrees of freedom for PDP modeling can be gained by including an additional within-subject manipulation that leaves the controlled process parameter C unaffected. Assuming that pay-

off manipulations for target objects (i.e., different rewards for correct *gun* and correct *tool* responses) affect response bias A selectively, this approach has previously been used by Klauer et al. (2015). Their results supported the discrimination bias hypothesis, that is, C varied in their study not only between primes but also between targets. To check the generalizability of their results, we manipulated response bias A selectively by employing different base-rates of gun targets in the WIT to achieve an identifiable generalized PDP procedure. The same manipulation has previously been used successfully for similar procedures such as the Implicit Association Test (Conrey et al., 2005). Like the pay-off manipulation, increasing the proportion of gun targets in the WIT is expected to leave the probability of correct target identification (i.e., C) unaffected, while it likely boosts overall response bias as reflected in the A-parameter of the PDP. Based on an extended WIT data structure that encompasses low and high base rates of gun targets, we can thus investigate whether (1) the racial bias effect is reflected in the prime-target interaction for the C-parameter as predicted by the discrimination bias hypothesis, (2) in the difference $A_B > A_W$ as predicted by the response bias hypothesis, or (3) both[1].

Research Overview

In the present research, we conducted two pre-registered studies to investigate discrimination bias and response bias as potential sources of racial bias in the WIT (Klauer et al., 2015; Payne et al., 2005). Study 1 made use of a base-rate manipulation for the target objects. Study 2 replicated Study 1. In addition, Study 2 investigated whether the instruction to make use of the prime face for target

[1] This analysis strategy deviates slightly from the pre-registered analysis plan. The pre-registered plan aimed to compare the goodness-of-fit of different restriction patterns that represent the two hypotheses of interest (discrimination bias vs. response bias, respectively; cf. Klauer et al., 2015). In contrast, the analysis procedure applied here investigated parameter estimates for the generalized PDP model directly to check for the third possibility that racial bias is mediated by response bias and discrimination bias simultaneously.

identification ("racial profiling instruction", cf. Payne et al., 2002) moderates the processes involved in the racial bias effect. We report for both studies how we determined our sample size, all data exclusions, all manipulations, and all measures. The pre-registrations are available at https://osf.io/4yxv9/?view_only=aa0ecf147152463ea657b24375f6c8f2.

Study 1:

Study 1 consisted of two within-subject conditions of the WIT with different base-rates of gun and tool targets. First, we investigated whether racial bias in weapon identification can be observed in error rates (Payne, 2001; Rivers, 2017). Second, we used a generalized PDP model to determine which mechanism can best account for the *C*- and *A*-parameter differences across prime and target conditions.

Method

Participants. We aimed for a sample size of 170 participants to match the sample size used in Study 5 of Klauer et al. (2015) who performed a similar model comparison for the WIT. To achieve this sample size and control for participants performance in an online study, we continued data collection until our target sample size was reached after excluding participants according to the pre-registered exclusion criteria. Specifically, participants were excluded when they failed in the attention check test and when their percentage of missing responses was an extreme outlier according to Tukey's criterion (i.e., 1.5 times the interquartile range above the median, cf. Clark-Carter, 2004, Chapter 9; Klauer et al., 2015). The collected sample consisted of 211 participants (Missing responses: *Mean* = 15.9%, *SD* = 27.8%, *Median* = 2.8%). We excluded 40 participants with more than 27.8% of missing responses, resulting in a final sample size of $N = 171$. Mean age was 24.1 years ($SD = 7.0$) and participants indicated their gender as 128 female, 36 male, and 7 other. Self-reported race of participants was 7.0% Black, 61.4% White, and 31.6% other.

Design. Independent variables were the base-rate (more guns vs. more tools), the prime face (Black vs. White) and the target object (gun vs. tool). All three independent variables were completely cross-classified in a balanced 2 x 2 x 2 within-subject design. The participants' error rates

per condition (percentage of incorrect "*gun*" or "*tool*" responses, depending on trial type) served as the dependent variable.

Materials. Prime faces, target objects, and the pattern mask were taken from Rivers (2017). Targets were five drawings of hand tools and five drawings of weapons presented in four different orientations rotated by 90 degrees. Target objects and the pattern mask can be accessed via the Open Science Framework (OSF): https://osf.io/9e6sa/. In total, prime faces consisted of 24 Black, 24 White male faces with neutral facial expression, and one neutral prime consisting of the outline of a face which was only used for the practice trials. Prime faces were provided by the first author of Phills et al. (2011). Each image was displayed with a size of 300 x 300 pixels.

Procedure. Data collection was conducted online via Prolific (www.prolific.co) [2021]. Participants were required to be at least 18 years of age and be part of the Prolific US-sample. The study took about 25 minutes and participants were rewarded with 3.20£ or equivalent. After providing consent, participants completed the WIT.

The WIT consisted of 24 practice trials, followed by 200 experimental trials for each base-rate condition. Trials were assigned to each participant in random order. The More Tools (MT) condition consisted of a 70%:30% ratio of tools relative to guns. Correspondingly, the More Guns (MG) condition consisted of a 30%:70% ratio of tools relative to guns. Participants were informed about the percentage of tool and gun targets beforehand. As an attention check, participants were instructed to type the percentage of targets they will see in an open text field on the following screen. To rule out order effects, the base-rate conditions were presented in random order for each participant.

In each trial, participants saw a sequence of a fixation cross (500 ms), a face prime (200 ms), a target object (200 ms), a pattern mask (300 ms), and a feedback screen (1000 ms) in the center of the screen. In practice trials, possible feedbacks were "correct", "false", or "too slow". In experimental trials, either no feedback occurred or "too slow" was provided as feedback. Participants had a response time limit of 500 ms after target onset. Participants were instructed to identify as fast and accurately as possible the target object while ignoring the face prime. For each participant, responding with "gun" or "tool" was randomly assigned to response keys *D* and *L*. After completing the WIT,

participants completed a basic demographics questionnaire (age, gender, and race) and were asked what they thought the study was about. At the end of the study, participants were thanked and debriefed.

MPT Modeling Procedure. We used the Bayesian hierarchical latent-trait MPT approach of Klauer (2010) as implemented in TreeBUGS (Heck et al., 2018) to estimate the PDP. For assessing goodness-of-fit, we used the test statistics T_1 and T_2 to calculate Bayesian posterior predictive p-values. T_1 summarizes how well the model accounts for the average response frequencies across participants. T_2 summarizes how well the model accounts for the variances and correlations of the response frequencies across participants. The posterior predictive p-value represents the comparison of the calculated test statistics obtained for observed and predicted response frequencies, with a value of $p > .05$ typically seen as evidence that model assumptions are in line with the data (Klauer, 2010; Klauer et al., 2015).

For Bayesian model comparison, we used the Deviance Information Criterion (DIC; Klauer et al., 2015; Spiegelhalter et al., 2002) and the Widely Applicable Information Criterion (WAIC; Vehtari et al., 2017; Watanabe, 2010) as information criteria taking model complexity into account. The model with the lowest DIC (WAIC) value represents the best compromise between model fit and model complexity. An ΔDIC (ΔWAIC) difference larger than two is typically interpreted as evidence for one model compared to the other (Burnham & Anderson, 2004; Klauer et al., 2015; Spiegelhalter et al., 2002). We used the R package TreeBUGS (Heck et al., 2018) to estimate the MPT-models of interest. The Markov Chain Monte Carlo (MCMC) algorithm was run for three independent estimation chains with 1,000,000 iterations each, of which 100,000 were removed as a burn-in. Every 500th iteration was retained to compute summary statistics. The Rubin-Gelman statistics $\hat{R}$ was smaller than 1.05 for all parameter estimates across all models, showing an acceptable convergence of MCMC sampling.

Results

Prior to all analyses, we excluded trials with latencies <100 ms and >1500 ms, resulting in exclusion of 4.76% of the trials. Given $N = 171$ participants, a sensitivity power analysis revealed that

a relatively small within-subject population effect size of $\boldsymbol{\eta_p^2} = .046$ (cf. Cohen, 1988)[2] can be detected in repeated measures ANOVA $F(1,170)$ tests with a Type-1 error level of $\alpha = .05$ and a statistical power of $1\text{-}\beta = .80$ (Faul et al., 2009). This holds for any main or interaction $F(1, 170)$ test (and simple main effect t-test) in Study 1. Thus, our study is sufficiently powered.

ANOVA Analyses. A 2 (prime) x 2 (target) repeated analysis of variance (ANOVA) of the error rates (Table 1) pooled across base-rate conditions resulted in a significant prime-target interaction, $F(1,170) = 11.08$, $p = .001$, $\eta_p^2 = .061$, descriptively in line with the pattern indicative of racial bias. The three-way 2 (base-rate) x 2 (prime) x 2 (target) interaction was not significant, $F(1,170) = 1.77$, $p = .19$, $\eta_p^2 = .010$, indicating that the base-rate condition did not modulate the prime-target interaction.

Insert Table 1 about here

To take a closer look at the prime-target interaction, we calculated this interaction separately for each base-rate condition. In the MG condition, the prime-target interaction was significant $F(1,170) = 10.09$, $p = .002$, $\eta_p^2 = .056$. Guns were less often misidentified as tools after Black primes ($M = 16.0\%$, $SD = 10.2$) versus White primes ($M = 18.1\%$, $SD = 10.2$; $t(170) = 2.63$, $p = .009$, $d_Z = 0.201$), whereas tools were more often misidentified as guns after Black primes ($M = 54.1\%$, $SD = 20.3$) versus White primes ($M = 50.8\%$, $SD = 18.6$; $t(170) = 2.90$, $p = .004$, $d_Z = 0.221$). In the MT condition, the prime-target interaction was also significant $F(1,170) = 5.26$, $p = .023$, $\eta_p^2 = .030$. However, in this condition guns were not misidentified significantly less often as tools after Black primes ($M = 49.9\%$, $SD = 19.8$) versus White primes ($M = 51.3\%$, $SD = 18.3$; $t(170) = 1.29$, $p = .20$,

---

[2] In line with Cohen (1988), boldface notation indicates population effect sizes (e.g., $\boldsymbol{\eta_p^2}$) whereas standard notation indicates estimated sample effect sizes (e.g., η_p^2) in what follows.

$d_Z = 0.098$), whereas tools were misidentified more often as guns after Black primes ($M = 16.5\%$, $SD = 9.9$) versus White primes ($M = 50.8\%$, $SD = 18.6$; $t(170) = 2.78$, $p = .006$, $d_Z = 0.212$).

MPT Modeling. The unconstrained PDP model showed acceptable model fit ($p(T_1) = .492$; $p(T_2) = .081$; Table 2). Mean parameter estimates and their corresponding 95%-Bayesian Credibility Interval (BCI) are listed in Table 3. To assess the influence of race on response bias, we calculated the ratio A_B/A_W which indicates racial bias against Black males when it is larger than one. The ratio was larger than one for the MG ($A_B/A_W = 1.059$ [1.020 – 1.103]) and the MT ($A_B/A_W = 1.187$ [1.061 – 1.326]) condition. To test whether these ratios differ significantly from one, we compared model fit, DIC and WAIC of the unconstrained model with models including the constraint $A_B = A_W$ for the respective base rate condition. The unconstrained model performed better (DIC = 7244.1, WAIC = 7063.4) than the respective $A_B = A_W$ constrained model for both the MG ($p(T_1) = .065$, $p(T_2) < .001$, DIC = 7381.2, WAIC = 7303.2) and the MT ($p(T_1) = .154$, $p(T_2) = .006$, DIC = 7290.4, WAIC = 7170.4) condition, suggesting an influence of race on response bias.

To investigate the influence of race on object discrimination, we calculated the ratios C_{BG}/C_{BT} and C_{WG}/C_{WT} which indicate how Black vs. White primes, respectively, facilitate discrimination of the gun targets relative to tool targets. The ratios C_{BG}/C_{BT} (1.544 [1.167 – 2.078]) and C_{WG}/C_{WT} (1.470 [1.127 – 1.956]) were both larger than one. To test whether the ratios C_{BG}/C_{BT} and C_{WG}/C_{WT} differ significantly from each other in an ordinal fashion, we compared model fit of a baseline model assuming $C_{B,T} = s_B \cdot C_{B,G}$ and $C_{WT} = s_W \cdot C_{W,G}$ to a model with the same parameter structure and the additional equality constraint $s_B = s_W$. This comparison allows to test for an ordinal interaction effect of prime and target condition for the controlled process parameter (Kuhlmann et al., 2019). This test revealed that the model with the equality constraint $s_B = s_W$ fitted the data better ($p(T_1) = .117$, $p(T_2) = .011$, DIC = 7287.1, WAIC = 7146.0) than the unconstrained model ($p(T_1) = .118$, $p(T_2) = .009$, DIC = 7302.9, WAIC = 7180.5). Thus, there is no support for the discrimination bias hypothesis according to which racial bias is mediated by object discrimination. A model constraining the shrinkage parameters equal to one (i.e., $s_B = s_W = 1$) did not fit the data well ($p(T_1) = .001$, $p(T_2) < .001$, DIC = 7425.3, WAIC = 7398.4), indicating that there is a higher probability of

successful controlled responding for gun compared to tool targets. Similarly, a model constraining all controlled process parameters to be equal (i.e., $C_{B,G} = C_{B,T} = C_{W,G} = C_{W,T}$) did not fit the data well ($p(T_1) = .001$, $p(T_2) < .001$, DIC = 7418.8, WAIC = 7395.2).

Insert Table 2 & 3 about here

Discussion

Across both base-rate conditions, participants displayed a weapon identification bias modulated by prime race. Tool targets were more often misidentified as guns after Black male faces compared to White male faces. Furthermore, results of Study 1 supported the hypothesis that racial bias in the WIT is mediated by response bias but did not support discrimination bias as a mechanism driving racial bias in the WIT. Our results rest on the assumption that base-rate manipulations of target objects selectively affect response bias while leaving discrimination unaffected. Notably, this is a standard assumption often made in discrimination and recognition research (e.g., Buchner et al., 1995; Conrey et al., 2005).

Although Study 1 suggests that racial bias in the WIT is more likely mediated by response bias than by discrimination processes, it is silent about the generalizability of this finding. For example, it remains an open question so far whether the finding still holds when participants are prompted to use race information as a cue to identify the following target object. We therefore investigated this research question in Study 2.

Study 2

Study 2 included two experimental groups that differed in task instructions. The control group of Study 2 was a direct replication of Study 1. The "racial profiling" group received an additional instruction to use the race of the prime for object discrimination (as detailed below). Hence, Study 2 allowed us to investigate the replicability of Study 1 on the one hand and the generalizability of results to participants working under "racial profiling" instructions on the other hand.

Method

Participants. As in Study 1, we aimed at a sample size of 170 participants for each of the two experimental groups. We used the same data collection strategy as in Study 1, including an attention check test for the racial profiling group. The collected sample consisted of 430 participants (Missing responses: *Mean* = 8.9%, *SD* = 15.8%, *Median* = 2.3%). We excluded 88 participants with more than 18.3% missing responses, resulting in a final sample size of $N = 171$ for the control group and $N = 171$ for the racial profiling group. Mean age was 35.3 years (SD = 12.6) and participants indicated their gender as 170 female, 164 male, and 8 other. Race of participants was 9.4% Black, 68.1% White, and 22.5% other.

Design. Independent variables were group (experimental vs. control, between Ss), base rate (more guns vs. more tools, within S), prime face (Black vs. White, within S) and target object (gun vs. tool, within S). All four independent variables were completely cross-classified in a balanced 2 x 2 x 2 x 2 mixed design. The participants' error rates per condition served as the dependent variable.

Materials. Target images were the same as in Study 1. The racial profiling instruction employed in the experimental group was taken from Payne et al.'s (2002) "Use Race" instruction:

> "The faces will be from either White (European American) or Black (African American) people. Research has shown that the race of the face sometimes impacts the ways that people classify the second object. People are sometimes faster and more accurate in responding to guns after a Black face than after a White face. You have been randomly assigned to take the "racial profiling" condition. Regardless of your personal views, we would like you to play the role of someone engaged in racial profiling. That is, try to make correct classifications, but we would like you to use the race of the faces to help you identify the gun or tool in question."

Procedure. Data collection was conducted online via Prolific with the same sample specifications as in Study 1. The study took about 28 minutes and participants were rewarded 3.50£. The study procedure was the same as in Study 1 except that the racial profiling group received the additional racial profiling instruction after they provided their consent. As additional attention check

test in this condition, participants had to indicate the experimental group they have been assigned to in an open text field presented on the screen that followed the racial profiling instruction.

MPT Modeling Procedure. The modeling procedure was the same as in Study 1. Models were estimated separately for the control and racial profiling group. The Rubin-Gelman statistics $\hat{R}$ was smaller than 1.05 for all parameter estimates across all models, showing an acceptable convergence of MCMC sampling.

Results

Prior to all analyses, we excluded trials with latencies <100 ms and >1500 ms resulting in the exclusion of 3.6% of the trials. Within each of the two experimental groups of participants, statistical power of the $F(1, 170)$ tests (and simple main effects t-tests) is identical to Study 1. For the $F(1, 340)$ tests involving both experimental groups, statistical power is even higher. Thus, like Study 1, Study 2 is well-powered.

ANOVA Analyses. A 2 (prime) x 2 (target) ANOVA for the error rates resulted in significant interactions for both the control group, $F(1, 170) = 33.85$, $p < .001$, $\eta_p^2 = .166$, and the racial profiling group, $F(1, 170) = 51.66$, $p < .001$, $\eta_p^2 = .233$. Replicating Experiment 1, these interactions are in line with the expected weapon identification bias. The three-way 2 (base-rate) x 2 (prime) x 2 (target) interaction was significant for the control group, $F(1, 170) = 7.25$, $p = .008$, $\eta_p^2 = .041$ and just insignificant for the racial profiling group, $F(1, 170) = 3.69$, $p = .056$, $\eta_p^2 = .021$, indicating that the base-rate condition moderated the size of the prime-target interaction at least in the control condition, in contrast to what we observed in Study 1. The three-way 2 (instruction) x 2 (prime) x 2 (target) between-group interaction was also significant, $F(1, 340) = 23.00$, $p < .001$, $\eta_p^2 = .063$. This result indicates a stronger prime-target interaction in the racial profiling group compared to the control group.

To take a closer look at the prime-target interaction, we also calculated these interactions separately for each base-rate and instruction condition. Table 1 lists mean error rates and their standard deviations for each instruction x base-rate x prime x target combination. For the control group, the respective prime-target interaction was significant for both the MG condition, $F(1,170) =$

5.69, $p = .018$, $\eta_p^2 = .032$, and the MT condition , $F(1,170) = 30.31$, $p < .001$, $\eta_p^2 = .151$. Guns were misidentified less often as tools after Black primes in both the MG, $t(170) = 2.54$, $p = .012$, $d_Z = 0.194$ and the MT condition $t(170) = 4.03$, $p < .001$, $d_Z = 0.308$. Also, tools were misidentified more often as guns after Black primes versus White primes. Whereas this effect just failed to be significant in the MG condition, $t(170) = 1.67$, $p = .097$, $d_Z = 0.128$, it was significant in the MT condition $t(170) = 5.16$, $p < .001$, $d_Z = 0.395$. For the racial profiling group, the respective prime by target interaction was significant for the MG, $F(1,170) = 42.04$, $p < .001$, $\eta_p^2 = .198$, and the MT condition , $F(1,170) = 55.49$, $p < .001$, $\eta_p^2 = .246$. Guns were misidentified less often as tools after Black primes in both the MG condition, $t(170) = 6.06$, $p < .001$, $d_Z = 0.464$, and the MT condition $t(170) = 8.13$, $p < .001$, $d_Z = 0.622$, whereas tools were misidentified more often as guns after Black primes in the MG condition, $t(170) = 5.94$, $p < .001$, $d_Z = 0.454$, and the MT condition $t(170) = 5.51$, $p < .001$, $d_Z = 0.421$.

MPT Modeling. The generalized PDP model showed acceptable fit for the control ($p(T_1) = .462$; $p(T_2) = .119$) and the racial profiling group ($p(T_1) = .520$; $p(T_2) = .291$; Table 2). Mean parameter estimates and their corresponding 95%-Bayesian Credibility Intervals (BCI) are listed in Table 3. Regarding the influence of race on response bias, the ratio A_B/A_W was larger than one in the control group for both the MG (1.027 [0.993 – 1.063]) and the MT (1.280 [1.141 – 1.432) condition. For the racial profiling group, A_B/A_W was also larger than one for the MG (1.078 [1.027 – 1.128]) and the MT (1.352 [1.155 – 1.564) condition. To test whether $A_B/A_W = 1$ holds, we compared model fit, DIC, and WAIC between the unconstrained models and the models including the constraint $A_B = A_W$, separately for the respective experimental group and base-rate condition. For the control group, the unconstrained model performed better (DIC = 7122.9, WAIC = 6960.6) than the model with the $A_B = A_W$ constraint for the MT condition ($p(T_1) < .001$, $p(T_2) = .035$, DIC = 7180.4, WAIC = 7057.9). For the MG condition, the result was similar for WAIC but inconclusive for DIC ($p(T_1) = .385$, $p(T_2) = .107$, DIC = 7123.0, WAIC = 6980.5), indicating partial support for the influence of race on response bias. For the racial profiling group, the unconstrained model performed better (DIC = 7201.8, WAIC = 6958.2) than the model with the $A_B = A_W$ constraint for both the MG ($p(T_1) <$

.001, $p(T_2) < .001$, DIC = 7472.9, WAIC = 7421.1) and the MT condition ($p(T_1) < .001$, $p(T_2) <$ .001, DIC = 7535.1, WAIC = 7467.1), also indicating support for the response bias hypothesis.

Regarding the influence of race on object discrimination in the control group, the ratios $C_{B,G}/C_{B,T}$ (1.377 [1.109 – 1.735]) and $C_{W,G}/C_{W,T}$ (1.278 [1.047 – 1.576]) were both larger than one. In the racial profiling group, the ratio $C_{B,G}/C_{B,T}$ was also larger than one (1.725 [1.348 – 2.233]) while $C_{W,G}/C_{W,T}$ (0.928 [0.741 – 1.144]) was smaller than one. To test whether these ratios differ significantly from each other, we again compared the order constrained base model assuming $C_{B,T} = s_B \cdot C_{B,G}$ and $C_{WT} = s_W \cdot C_{W,G}$ with a model that additionally equates $s_B = s_W$.

For the control group, this comparison revealed that the equated model ($p(T_1) = .256$, $p(T_2) = .027$, DIC = 7172.8, WAIC = 7065.3) fitted the data better than the unconstrained model ($p(T_1) = .226$, $p(T_2) = .026$, DIC = 7178.5, WAIC = 7075.5), indicating no support that racial primes affect object discrimination. A model assuming $s_B = s_W = 1$ did not fit the data well ($p(T_1) < .001$, $p(T_2) = .002$), indicating more successful controlled responding for gun compared to tool targets. Similarly, a model constraining all controlled process parameters to be equal (i.e., $C_{B,G} = C_{B,T} = C_{W,G} = C_{W,T}$) did not fit the data well ($p(T_1) < .001$, $p(T_2) = .003$, DIC = 7297.1, WAIC = 7278.6).

For the racial profiling group, the order-constrained base model assuming $C_{B,T} = s_B \cdot C_{B,G}$ and $C_{W,T} = s_W \cdot C_{W,G}$ did not fit the data ($p(T_1) = .026$, $p(T_2) = .123$). This reflects the fact that $C_{W,T} = s_W \cdot C_{W,G}$ cannot account for the data because $C_{W,G}$ is smaller than $C_{W,T}$. The misfit of this model thus provides evidence for a disordinal prime-target interaction in the C-parameter, showing that racial bias is mediated by target discrimination in the experimental group. Accordingly, a model that constrains all controlled process parameters to be equal (i.e., $C_{B,G} = C_{B,T} = C_{W,G} = C_{W,T}$) did not fit the data well either ($p(T_1) < .001$, $p(T_2) = .022$, DIC = 7294.8, WAIC = 7180.3).

Discussion

A weapon identification bias modulated by race was observed across both base-rate conditions for both experimental groups. However, in line with Payne et al. (2002), the racial profiling group displayed stronger racial bias effects than the control group. Furthermore, the modeling results for the control group successfully replicate Study 1 and again supported the hypothesis that WIT

performance is driven by response bias rather than by discrimination bias. In contrast, the results for the racial profiling group indicated that both response bias and discrimination bias may be involved when a person is prompted to employ race primes for object identification. In sum, the results of Study 2 replicate Study 1 by showing that racial bias in the WIT is due to response bias in the first place. However, when participants are explicitly instructed to make use of the prime's race for object identification, discrimination bias may boost racial bias in addition.

General Discussion

The current research examined the source of racial bias as typically observed in the Weapon Identification Task (WIT). Overall, results suggest a dominant role of response bias, a conclusion that is congruent with previous WIT research (Payne et al., 2005). We also found that discrimination bias may add to racial bias in the WIT (Klauer et al., 2015). However, in our research, this happened only when participants were explicitly instructed to use the race of the prime for target identification.

Strong Evidence for Response Bias underlying WIT performance

The results of both of our studies indicate that response bias is a major determinant of racially biased responding in the WIT. This means that automatic stereotypic associations elicited by the prime create an additional stream of information that induces response bias moderated by race whenever controlled processing fails. This aligns well with other model-based analyses that observed response biases moderated by primes' race. For example, according to the diffusion model analysis of Todd et al. (2021), participants have a higher initial bias towards gun responses following Black faces, a preference that is stronger when race information is salient. Laukenmann et al. (2023) used response time-extended MPT models to compare different process models based on the PDP. Their model comparison favored the Default Interventionist Model (DIM; Klauer & Voss, 2008) according to which an initial default response moderated by race primes is activated instantly, in line with the findings of Todd et al. (2021). In the framework of the DIM, response bias, mirrored in the A-parameter, only drives the response if controlled processing fails to discriminate the target object correctly or to resolve possible response conflict if the default response is incongruent with the target (Klauer & Voss, 2008; Laukenmann et al., 2023).

In line with this, several previous findings suggest that racial bias depends on participants' cognitive and motivational abilities to exert control. For example, people motivated to control their prejudices (Volpert-Esmond et al., 2020) or individuals with higher inhibitory capacities (Ito et al., 2015) display overall better WIT performance. Also, participants with depleted executive resources show more racial bias (Govorun & Payne, 2006). Hence, our results align well with previous research on response bias as a major mechanism driving racial bias in the WIT. Specifically, automatic default responses may lead to weapon identification bias if these default responses are not corrected effectively by controlled processes.

Mixed Evidence for Discrimination Bias in the WIT

The results of Study 2 suggest that discrimination bias can additionally boost racial bias if participants are instructed to use the primes' race as a cue for responding. Hence, if participants are inclined to use the primes' race for object discrimination, this may influence the interpretation of early chunks of information in the light of contextual cues provided by the racial stereotype. Consequently, when racial information is salient, participants might misperceive presented objects and create distorted perceptions, such as confusing the shaft of a screwdriver with a barrel of a gun after seeing a Black face prime (Klauer et al., 2015; Klauer & Voss, 2008; Payne et al., 2005).

Addressing a similar research question, Todd et al. (2021) manipulated race salience in the WIT by asking participants to sort faces by race (Race-salient condition: Black vs. White faces) or by another, non-racial criterion like age (Age-salient condition: Young vs. Old faces). In their diffusion model analysis of the WIT which participants performed afterwards, race salience only led to a generally faster rate of object information extraction following Black faces. However, no prime-target interaction effect on information extraction was observed, irrespective of whether the race was salient or not. This suggests that race salience does not moderate discrimination bias in the WIT, in contrast to our results. A possible explanation is that the race sorting task used by Todd et al. (2021) generally boosts object discrimination following Black faces, whereas the racial profiling instruction we used in Study 2 explicitly ties the Black face with a better discrimination for weapon than for tool targets.

Hence, the nature of the race salience manipulation may influence whether race salience fosters discrimination bias in the WIT or not.

In contrast to both Todd et al.'s (2021) and our results, Klauer et al. (2015) found convincing evidence for discrimination bias even when race was not made salient. Notably, they used a German student sample and a high number of task trials in their Study 5. One possible explanation in line with our findings could be that race salience is generally increased for German university students because Black people are relatively rare in Germany compared to the United States. Also, a high number of task trials may motivate participants to rely more on the primes' race as additional information for object discrimination. Taken together, these differences in sample and study characteristics might have led participants to rely more on racial profiling as a response strategy.

In principle, differences in data analysis strategies employed by Klauer et al. (2015) and in our current research may affect the conclusions. In the analyses reported in the Results section, we assessed the effect of race on the C- and A-parameters of the generalized PDP model. This allows us to investigate whether (1) response bias, (2) discrimination bias or (3) both biases are involved in WIT performance. In contrast, Klauer et al. (2015) compared the DIC model selection statistics of three generalized PDP models: model M1 with the C-parameters allowed to vary between target types and prime races (i.e., the model we used), model M2a with the C-parameters equated across target types (i.e., the conventional PDP specification), and M2b with the C-parameter equated across prime races. They report a better fit for M1 compared to M2a and interpreted this as support for a prime-target interaction effect of the C-parameter. However, they also report a better fit of M2b compared to M1. This suggests that there is a strong influence of target type on the C-parameter while the effect of the prime on the C-parameter is negligible. Note that the same pattern is also evident in our results of Study 1 and in the control group of Study 2, as the C-parameters are reliably larger for gun compared to tool targets in both studies. Thus, despite different data analysis strategies, Klauer et al.'s and our current work exhibit more similarities of results than the differences in conclusions concerning the role of discrimination bias in the WIT might suggest.

A recent study by Stein et al. (2023) approached the possible role of discrimination bias underlying the WIT in a different way. Stein et al. (2023) investigated whether participants become aware of weapons earlier if preceded by a Black male face. Hence, their study did not focus on whether race distorts the early perception and interpretation of an object, but rather whether objects have a lower stereotype-congruent awareness threshold in the first place (i.e., a faster awareness of guns compared to tools following Black face prime). In addition to a standard WIT, their participants performed a breaking Continuous Flash Suppression task (b-CFS). Using a stereoscope, the b-CFS is a binocular rivalry task with the target object presented to one eye and a CFS flashing mask to the other eye, preceded by a face prime for both eyes. In an experimental trial, the target objects' transparency was decreased whereas CFS masks' transparency was increased. Participants response time to locate the target object above or below a fixation cross is interpreted as measure for when the target object enters participants' awareness. Although Stein et al. (2023) replicated a typical weapon identification bias in the WIT, the b-CFS showed no corresponding bias. They concluded that weapon identification bias does not result from initial perceptual processes but response-related processes. This converges with our findings in Study 1 and in the control group of Study 2.

The Role of Conflict Resolution

As mentioned before, besides their important role for correct object perception and interpretation, controlled processes may also be important for resolving conflicts between different streams of task information (Ito et al., 2015; Klauer et al., 2008; Laukenmann et al., 2023). Hence, these conflict resolution processes might provide an alternative explanation how object discrimination may vary as a function of prime and target, in contrast to biased early perceptions and interpretations of target objects.

In the conflict resolution account, participants are more successful in integrating the extracted perceptual information from the target object and the default response elicited by the prime race if both pieces of information are congruent rather than incongruent. For example, a gun target following a Black face prime produces less conflict between the default gun response and the perceptual information compared to a tool target following a Black face prime. Although Study 1 and the control

group in Study 2 showed no prime-target interaction effect on object discrimination, the racial profiling group in Study 2 did. The conflict resolution account would explain this by an increased cognitive conflict when race is salient. The reason is that the racial profiling instruction makes it harder for participants to inhibit salient incongruent default responses, resulting in worse performance in incongruent trials and better performance in congruent trials.

Unfortunately, our generalized PDP model does not provide for a straightforward way to disentangle whether early perceptual processes, conflict resolution, or both underly discrimination bias as observed for the racial profiling group in Study 2. Hence, this remains an open question for future research.

Specification of the Process Dissociation Model

Our research is also informative with respect to the question whether the PDP in its conventionally used form – equating the controlled process parameter across target conditions – is a reasonable measurement model to estimate the influence of automatic and controlled processes on WIT performance (Payne, 2001). Employing a generalized PDP model for two base-rate conditions, we observed no interaction effect on the *C*-parameter in Study 1 and in the control group of Study 2. However, a main effect of the target object was observed in both studies, with reliably larger *C*-estimates for gun than for tool targets. This aligns well with the finding of Klauer et al. (2015) that equating the *C*-parameter across primes (resulting in one *C*-parameter for guns and another for tools), may also be a legitimate generalized PDP model specification besides the conventional PDP model. Thus, when only the standard four conditions are available (Black versus White primes crossed with guns versus tools), the PDP model best supported by the current evidence would include two separate *C*-parameters for guns and tools next to two separate *A*-parameters for Black and White face primes. To be on the safe side, however, it may be advantageous to use a generalized PDP model in future research that allows the *C*-parameter to vary freely between all prime race and target type conditions (i.e., model M1 of Klauer et al., 2015, which is also the model we used). As Study 2 shows, this is particularly important when experimental manipulations like the racial profiling instruction are used that might foster discrimination bias in addition to response bias.

Limitations

Two possible limitations of our current research should be discussed briefly. First, the findings of the current study rely on the assumption that the *C*-parameter is not affected by the base-rates of targets. Notably, this is a common assumption in discrimination and recognition research (Buchner et al., 1995; Conrey et al., 2005) that has often been used in different contexts and does not appear to be problematic. In this sense, our procedure is comparable to the pay-off manipulation used by Klauer et al. (2015) to manipulate response bias *A* selectively, also based on the assumption that *C* is not affected by pay-offs.

Second, our base-rate manipulation led to higher error rates for low base-rate targets compared to high base-rate target. This pattern is consistent across both studies and independent of target object type. Importantly, however, despite this difference in accuracy for low and high base rate targets, the error rates reflected a prime-target interaction indicative of racial bias in all studies. Regarding PDP-parameter estimates, the base-rate manipulation is mirrored in the size of *A*-parameters as intended, that is, with large *A*-parameters in the more gun base-rate condition and small *A*-parameters in the more tool base rate condition. *C*-parameter estimates generally entailed small values. However, *C* estimates were clearly larger than zero which shows that controlled responding occurred in our studies. In sum, although error rates tended to be high in the low base-rate conditions, this does not call into question that the generalized PDP can be used to assess the role of response bias and discrimination bias as mechanisms driving racial bias in the WIT.

Conclusion

In two independent pre-registered studies based on $N = 513$ participants, we found support for the hypothesis that response bias is the main mechanism driving racial bias in the WIT. This means that automatic stereotypic associations, elicited by the prime's race, lead to an additional, interfering stream of information besides controlled discrimination processes. This may lead to the typical racially biased response patterns in weapon identification when controlled object discrimination fails. This notwithstanding, if participants attend to race during object identification

(e.g., because race information is salient), discrimination bias may play an additional role in racially biased responding. Specifically, the primes' race may provide stereotype-biased context cues when participants aim to resolve the perceptual ambiguity in target object identification. In sum, both biases can lead to racially biased responding in weapon identification, although discrimination bias becomes important only when people attend to race as a task-relevant cue. From an applied perspective, the two biases suggest at least two strategies to weaken racial biases in spontaneous behavior: First, reducing response bias effects by enhancing the capacity to inhibit automatic threat-based associations and second, facilitating unbiased object discriminations by reducing the salience of race.

References

Bishara, A. J., & Payne, B. K. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology*, *45*(3), 524–534. https://doi.org/10.1016/j.jesp.2008.11.002

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124*(2), 137–160. https://doi.org/10.1037/0096-3445.124.2.137

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Clark-Carter, D. (2004). *Quantitative Psychological Research: A Student's Handbook*. New York, NY: Psychology Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum

Conrey, 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*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi:10.3758/BRM.41.4.1149

Govorun, O., & Payne, B. K. (2006). Ego-depletion and prejudice: Separating automatic and controlled components. *Social Cognition, 24*(2), 111–136. https://doi.org/10.1521/soco.2006.24.2.111

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

Huntsinger, J. R., Sinclair, S., & Clore, G. L. (2009). Affective regulation of implicitly measured stereotypes and attitudes: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *45*(3), 560–566. https://doi.org/10.1016/j.jesp.2009.01.007

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Klauer, K. C. (2010). Hierarchical Multinomial Processing Tree Models: A Latent-Trait Approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221. https://doi.org/10.1037/xge0000044

Klauer, K. C., & Voss, A. (2008). Effects of Race on Responses and Response Latencies in the Weapon Identification Task: A Test of Six Models. *Personality and Social Psychology Bulletin*, *34*(8), 1124–1140. https://doi.org/10.1177/0146167208318603

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*(4), 215–229. https://doi.org/10.1016/j.jmp.2004.03.002

Kuhlmann, B. G., Erdfelder, E., & Moshagen, M. (2019). Testing interactions in multinomial processing tree models. *Frontiers in Psychology*, *10*, 2364. https://doi.org/10.3389/fpsyg.2019.02364

Laukenmann, R., Erdfelder, E., Heck, D. W., & Moshagen, M. (2023). Cognitive processes underlying the weapon identification task: A comparison of models accounting for both response

frequencies and response times. *Social Cognition*, *41*(2), 137–164. https://doi.org/10.1521/soco.2023.41.2.137

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192. https://doi.org/10.1037/0022-3514.81.2.181

Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology*, *62*, 1–50. https://doi.org/10.1016/bs.aesp.2020.04.001

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology*, *38*(4), 384–396. https://doi.org/10.1016/S0022-1031(02)00006-9

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology*, *41*(1), 36–47. https://doi.org/10.1016/j.jesp.2004.05.001

Phills, C. E., Kawakami, K., Tabi, E., Nadolny, D., & Inzlicht, M. (2011). Mind the gap: Increasing associations between the self and blacks with approach behaviors. *Journal of Personality and Social Psychology, 100*(2), 197–210. https://doi.org/10.1037/a0022159

Rivers, A. M. (2017). The Weapons Identification Task: Recommendations for adequately powered research. *PLOS ONE*, *12*(6), e0177857. https://doi.org/10.1371/journal.pone.0177857

Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*. Advance Online Publication. https://doi.org/10.1037/met0000561

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stein, T., Ciorli, T., & Otten, M. (2023). Guns are not faster to enter awareness after seeing a Black face: Absence of race-priming in a gun/tool task during continuous flash suppression.

Personality and Social Psychology Bulletin, *49*(3), 405–414. https://doi.org/10.1177/01461672211067068

Todd, A. R., Johnson, D. J., Lassetter, B., Neel, R., Simpson, A. J., & Cesario, J. (2021). Category salience and racial bias in weapon identification: A diffusion modeling approach. *Journal of Personality and Social Psychology*, *120*(3), 672–693. https://doi.org/10.1037/pspi0000279

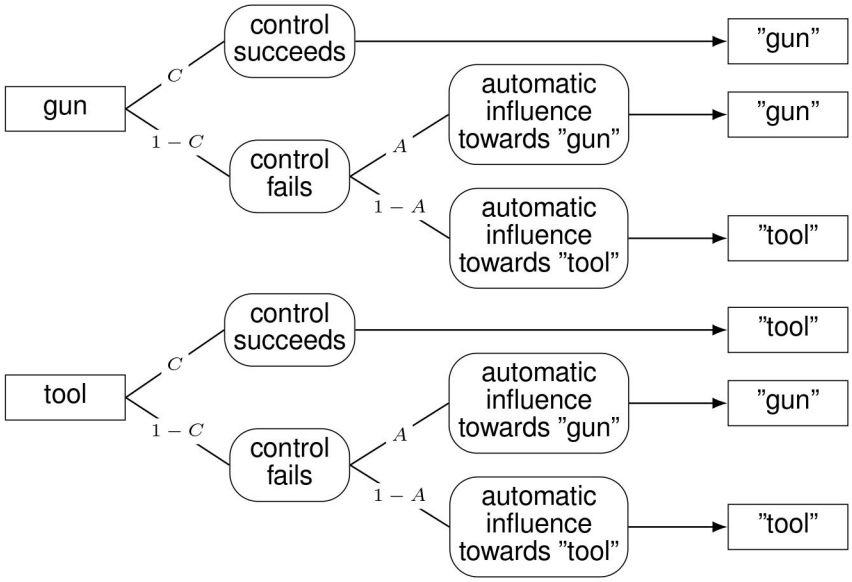
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Volpert-Esmond, H. I., Scherer, L. D., & Bartholow, B. D. (2020). Dissociating automatic associations: Comparing two implicit measurements of race bias. *European Journal of Social Psychology, 50*(4), 876–888. https://doi.org/10.1002/ejsp.2655

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Figure 1

Generic PDP model for the Weapon Identification Task.



Note. Process Dissociation Procedure (PDP; Figure taken from Laukenmann et al., 2023). Parameters C and A denote probabilities of response determination by a controlled process and an automatic process, respectively. Note that A is conditional on a failure of the controlled process, that is, A represents the conditional probability of response determination by an automatic process given controlled process failure.

Table 1

Mean proportion of errors by prime, target and base-rate condition in Study 1 and 2

	Prime			
	Black		White	
Target	*M*	*SD*	*M*	*SD*
Study 1				
More Gun condition				
Gun	.160	.102	.181	.102
Tool	.541	.203	.508	.186
More Tool condition				
Gun	.499	.198	.513	.183
Tool	.165	.099	.147	.100
Study 2				
Control Group				
More Gun condition				
Gun	.141	.100	.153	.100
Tool	.500	.204	.483	.203
More Tool condition				
Gun	.438	.192	.481	.205
Tool	.161	.111	.130	.099
Racial Profiling Group				
More Gun condition				
Gun	.131	.112	.196	.134
Tool	.532	.219	.448	.216
More Tool condition				
Gun	.404	.209	.524	.188
Tool	.205	.151	.140	.113

Table 2

Goodness-of fit statistics and model selection indices for Study 1 and 2

	Study 1				Study 2							
					Control				Racial Profiling			
Parameter constraints	$p(T_1)$	$p(T_2)$	DIC	WAIC	$p(T_1)$	$p(T_2)$	DIC	WAIC	$p(T_1)$	$p(T_2)$	DIC	WAIC
unconstrained	.492	.081	7244.1	7063.4	.462	.119	7122.9	6960.6	.520	.291	7201.8	6958.2
$A_B = A_W$ (More Guns)	.065	<.001	7381.2	7303.2	.385	.107	7123.0	6980.5	<.001	<.001	7472.9	7421.1
$A_B = A_W$ (More Tools)	.154	.006	7290.4	7170.4	<.001	.035	7180.4	7057.9	<.001	<.001	7535.1	7467.1
$C_{B,G} = C_{B,T} = C_{W,G} = C_{W,T}$	.001	<.001	7418.8	7395.2	<.001	.003	7297.1	7278.6	<.001	.022	7294.8	7180.3
$C_{B,T} = s_B \cdot C_{B,G}$; $C_{W,T} = s_W \cdot C_{W,G}$	.118	.009	7302.8	7180.5	.226	.026	7178.5	7075.5	.026	.123	7264.5	7086.2
$s_B = s_W$	.117	.011	7287.1	7146.0	.256	.027	7172.8	7065.3	<.001	.051	7249.1	7070.5
$s_B = s_W = 1$	.001	<.001	7425.3	7398.4	<.001	.002	7301.5	7283.1	<.001	.021	7299.4	7179.3

Note. C = controlled process, A = automatic process, s = shrinkage, B = Black prime, W = White prime, G = Gun target, T = Tool target.

Table 3

Parameter estimates for PDP model of Study 1 and 2

	Study 1		Study 2			
			Control group		Racial Profiling group	
Parameter	*Mean*	*95%-BCI*	*Mean*	*95%-BCI*	*Mean*	*95%-BCI*
$C_{B,G}$	.306	[.255 – .358]	.381	[.329 – .435]	.402	[.347 – .458]
$C_{B,T}$	.202	[.149 – .257]	.279	[.223 – .335]	.236	[.180 – .294]
$C_{W,G}$	.320	[.273 – .366]	.365	[.315 – .417]	.284	[.233 – .332]
$C_{W,T}$	.221	[.166 – .273]	.288	[.229 – .344]	.309	[.245 – .369]
$C_{B,G}/C_{B,T}$	1.544	[1.167 – 2.078]	1.377	[1.109 – 1.735]	1.725	[1.348 – 2.233]
$C_{W,G}/C_{W,T}$	1.470	[1.127 – 1.956]	1.278	[1.047 – 1.576]	0.928	[0.741 – 1.144]
C_{IA}	1.064	[0.752 – 1.463]	1.085	[0.830 – 1.384]	1.877	[1.373 – 2.549]
$A_{B,MG}$	.766	[.743 – .789]	.770	[.745 – .794]	.784	[.755 – .810]
$A_{W,MG}$	.723	[.696 – .749]	.750	[.723 – .776]	.728	[.699 – .756]
$A_{B,MT}$	.217	[.197 – .239]	.230	[.207 – .254]	.271	[.242 – .300]
$A_{W,MT}$	.183	[.164 – .203]	.180	[.162 – .200]	.201	[.179 – .223]
$A_{B,MG}/A_{W,MG}$	1.059	[1.020 – 1.103]	1.027	[0.993 – 1.063]	1.078	[1.027 – 1.128]
$A_{B,MT}/A_{W,MT}$	1.187	[1.061 – 1.326]	1.280	[1.141 – 1.432]	1.352	[1.155 – 1.564]

Note. C = controlled process, A = automatic process, B = Black prime, W = White prime, G = Gun target, T = Tool target, MG = More Guns base-rate condition, MT = More Tools base-rate condition. $C_{IA} = (C_{B,G}/C_{B,T})/(C_{W,G}/C_{W,T})$.

Cognitive Processes Underlying the Weapon Identification Task: A Comparison of Models Accounting for Both Response Frequencies and Response Times

Ruben Laukenmann[1], Edgar Erdfelder[1], Daniel W. Heck[2], and Morten Moshagen[3]

[1] University of Mannheim

[2] University of Marburg

[3] Ulm University

Author Note

Manuscript preparation was supported by the Research Training Group "Statistical Modeling in Psychology" (SMiP), funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, GRK 2277), and by a Doctoral Research Fellowship from the Stiftung der Deutschen Wirtschaft gGmbH (sdw, Foundation of German Business), funded by the German Federal Ministry of Education and Research (awarded to RL).

The authors are very grateful to M. J. Amon, A. J. Bishara, J. Correll, A. J. Lambert, C. Madurski, and A. M. Rivers for sharing the raw data we used in our model comparisons. We also thank Karl C. Klauer and an anonymous reviewer for helpful comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Ruben Laukenmann or Edgar Erdfelder, Cognition and Individual Differences Lab, A5, B 118, University of Mannheim, D-68159 Mannheim, Germany (email: ruben.laukenmann@psychologie.uni-mannheim.de or erdfelder@uni-mannheim.de).

Online supplementary material is available at: https://osf.io/7vjrq

We have no conflicts of interest to disclose.

Abstract

The *weapon identification task* (WIT) is a sequential priming paradigm designed to assess effects of racial priming on visual discrimination between weapons (guns) and innocuous objects (tools). We compare four process models that differ in their assumptions on the nature and interplay of cognitive processes underlying prime-related weapon-bias effects in the WIT. All four models are variants of the *process dissociation procedure*, a widely used measurement model to disentangle effects of controlled and automatic processes. We formalized these models as response time-extended multinomial processing tree models and applied them to eight data sets. Overall, the *default interventionist model* (DIM) and the *preemptive conflict-resolution model* (PCRM) provided good model fit. Both assume fast automatic and slow controlled process routes. Additional comparisons favored the former model. In line with the DIM, we thus conclude that automatically evoked stereotype associations interfere with correct object identification from the outset of each WIT trial.

Keywords: weapon identification task, process dissociation procedure, dual process models, multinomial processing tree modeling, racial bias

Word count: 9,394 (excluding references, figures, tables, and footnotes)

Cognitive Processes Underlying the Weapon Identification Task: A Comparison of Models Accounting for Both Response Frequencies and Response Times

Motivated by various incidents of police shootings of unarmed Black men, several studies investigated the visual discrimination between weapons and perceptually similar innocuous objects as a function of person characteristics such as race (Payne, 2001; Rivers, 2017). To investigate the influence of racial stereotypes on object identification, Payne (2001) proposed the weapon identification task (WIT). The WIT is a sequential priming paradigm. Participants are instructed to identify visually presented objects as either weapons (e.g., guns) or innocuous objects (e.g., tools). In the standard version of the WIT, target objects are preceded either by a Black or a White male face (Rivers, 2017). Many studies found that innocuous objects are more often erroneously misidentified as weapons following a brief presentation of a Black male face compared to a brief presentation of a White male face. In addition, participants identify a weapon typically faster after seeing a Black male face prime and an innocuous object faster after seeing a White male face prime. This result pattern, as reflected in error rates and response times, is known as *weapon identification bias* (Payne, 2001; Rivers, 2017; Todd et al., 2021). According to the meta-analysis of Rivers (2017), the weapon identification bias has a large effect size for error rates ($\eta^2 = .204$) and a medium effect size for response times ($\eta^2 = .106$).

To investigate the cognitive processes underlying the weapon identification bias, Payne (2001) applied the *process dissociation procedure* (PDP), a measurement model that decomposes task performance in effects of controlled and automatic processes. Whereas the controlled process represents the ability to identify the target object correctly, the automatic process reflects a response bias towards weapons versus innocuous objects. The typical finding is that prime race does not affect the controlled process but the automatic process, with Black male face primes leading to a larger response bias towards weapons compared to White male face primes (e.g., Huntsinger et al., 2009; Ito et al., 2015; Payne, 2001; Payne et al., 2002).

Although the PDP model allows to disentangle and measure the contributions of automatic and controlled influences on performance, it is agnostic with respect to the exact specification and temporal order of these cognitive processes. In fact, as outlined by Klauer and Voss (2008), the PDP model is consistent with four different cognitive process models for the WIT. Identifying the empirically most adequate of these process models is important as different models suggest different intervention strategies to reduce weapon identification bias. Knowledge about the underlying processes enables a deeper understanding of the weapon identification bias and thus provides a better foundation for racial bias awareness programs and for racial bias reduction interventions (e.g., Kahn & Martin, 2020; Klauer & Voss, 2008; Swencionis & Goff, 2017).

In this article, we first introduce the PDP and revisit the four psychological process models based on the PDP. We then formalize these process models in the framework of *response time-extended multinomial processing tree* (MPT-RT) models (Heck & Erdfelder, 2016) and compare them across eight different data sets of the weapon identification task. Including response latencies in the PDP via MPT-RT modeling allows us to compare the four models empirically and to identify the best-fitting process model. Hence, the current research goes beyond previous qualitative comparisons based on observed response times and accuracy patterns (Klauer & Voss, 2008) by providing quantitative assessments of the latent processing mechanisms underlying empirical data. Such an assessment is not possible when response frequencies are considered in isolation, which has been routine practice in PDP analyses of WIT performance so far. Finally, we discuss implications as well as remaining limitations of our model comparison.

Process Dissociation Procedure

To measure the contributions of automatic and controlled processes to WIT performance, Payne (2001) proposed the process dissociation procedure. Since then, the PDP model has become a widely used dual-process model for analyzing performance in the WIT (e.g., Bishara & Payne, 2009; Huntsinger et al., 2009; Ito et al., 2015; Klauer et al., 2015; Payne, 2005). According to this model, the controlled process represents the latent ability to identify the target and thus always leads to a correct response if it succeeds. The controlled process is thought to be effortful and constrained by

available cognitive resources (Klauer & Voss, 2008; Payne, 2001). In contrast, the automatic process reflects a latent response bias towards weapons rather than innocuous objects. This automatic bias is typically activated by a prime stimulus and driven by preexisting cognitive structures (e.g., stereotypical associations induced by a preceding face prime). The automatic process is assumed to be effortless and spontaneous (Conrey et al., 2005; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). Based on the participants' task performance, the PDP model allows researchers to estimate the probabilities that responses are determined by controlled or automatic processes. Parameters C and its complement $(1 - C)$ denote the probabilities that the controlled process succeeds or fails, respectively. Analogously, parameters A and $(1 - A)$ represent the probabilities that automatic bias drives the response towards *gun* or *tool*, respectively.

Figure 1 illustrates the PDP measurement model in the form of two processing tree diagrams, one for each target condition (i.e., gun and tool). The left side indicates the to-be-identified target object, followed by the possible processing branches (and their associated probabilities) that lead to certain responses in the WIT, as indicated on the right side of the trees. For the gun target tree (we refer to the weapon target condition as gun target condition), for example, the controlled process succeeds with probability C, resulting in the correct response *gun*. When the controlled process fails to determine the response with probability $(1 - C)$, the automatic process may trigger the response *gun* with probability A and the response *tool* with the complementary probability $(1 - A)$. For the tool target tree, the processing branches are analogous, except that the controlled process results in the correct response *tool*.

Insert Figure 1 about here

Parameters of processing tree models such as those illustrated in Figure 1 can be estimated using *multinomial processing tree* (MPT) modeling (for a tutorial, see Schmidt, Erdfelder, & Heck,

2023). To enable parameter estimation in MPT models, expected response probabilities are first derived by calculating the product of the probability parameters along each branch of a tree and then summing up those branch probabilities that refer to the same response. For example, a correct response in a gun target trial can be due to the controlled process being successful (with probability C) or due to an automatic response bias towards *gun* when the controlled process fails (with probability $(1 - C) \cdot A$). The probability of a correct response thus equals the sum of these two probabilities:

$$p(\text{response gun} \mid \text{target object gun}) = C + (1 - C) \cdot A$$

Based on a system of such model equations for all response categories in each condition, parameters are then estimated by minimizing an appropriate distance measure between observed and expected response frequencies (e.g., the log-likelihood ratio statistic G^2). MPT models have previously been used to estimate PDP parameters for the WIT (Bishara & Payne, 2009; Klauer et al. 2015). Compared to alternative estimation methods based on simple algebraic equations (Payne, 2001), MPT modeling enables not only conjoint estimation of all relevant PDP parameters along with their standard errors in one or more experimental conditions but, more importantly, also the assessment of model fit whenever the number of parameters is less than the number of independent response categories.

Nature of Cognitive Processes

Despite the clear-cut definition of the parameters C and A as probabilities of latent cognitive processes underlying responses in the WIT, their psychological nature has been a matter of debate (e.g., Bishara & Payne, 2009; Klauer & Voss, 2008; Klauer et al., 2015; Payne, 2005). Conrey et al. (2005) considered four potential cognitive determinants of performance in the WIT (see also Evans, 2007; Klauer & Voss, 2008; Sherman, 2006):

- the ability to discriminate between weapons and innocuous objects
- activation of racial associations (e.g., Black males are associated with guns)
- the ability to resolve conflicts between racial associations and target identification (the so-called overcoming bias)

- the guessing tendency towards one of the response options when none of the other processes determines the response

These four processes closely relate to controlled and automatic processes in the PDP. Specifically, both the ability to discriminate between stimuli and the overcoming bias reflect controlled processes, whereas activation of racial associations and guessing tendencies reflect automatic processes (Klauer & Voss, 2008). For the PDP this implies that parameter C can reflect target discrimination, overcoming bias, or both, whereas parameter A may reflect activation of racial associations, guessing tendencies, or both. In addition, given that controlled and automatic processes may operate sequentially or in parallel, it is worthwhile to define more precisely how these processes interact and, if so, how they are temporally ordered.

As the PDP is silent about the psychological nature of these processes and their temporal order and duration, several psychological process models based on the PDP are conceivable, each of which relies on different assumptions about the nature and the interplay of automatic and controlled processes.

Process models for the WIT

Based on Evans (2007), Klauer and Voss (2008) discussed four different psychological process models for the WIT that directly relate to the PDP. The *preemptive conflict-resolution model* (PCRM) assumes a preemptive decision whether the controlled process or the automatic process will determine the response. In comparison, the *default interventionist model* (DIM) presupposes that a default automatic response is always activated first which may or may not be successfully overcome by the controlled process subsequently. According to the *parallel competitive model* (PCM), in contrast, automatic and controlled processes run in parallel, sometimes requiring a subsequent conflict resolution process when the proposed responses disagree. Finally, the *guessing model* (GM) assumes that the controlled process determines the response, but if it fails to do so, the response is based on guessing.

As detailed below, all four psychological process models are based on different assumptions about the nature and the chronological order of the underlying cognitive processes. This

notwithstanding, they are all instantiations of the PDP that cannot be distinguished based on the accuracy of responses across the four prime-target conditions alone. In other words, these process models are empirically indistinguishable when formalized as simple PDP models for response frequencies.

However, these models come along with different assumptions about the relative latencies of the processing branches of the PDP (Klauer & Voss, 2008) and can thus be tested against each other when predictions concerning relative latencies are considered in addition. To investigate the relative latencies of process branches in MPT models, Heck and Erdfelder (2016) proposed *response time-extended multinomial processing tree* (MPT-RT) models. In the MPT-RT framework, responses are subclassified by their response time, for example, by a dichotomous split into fast versus slow responses. To transform an MPT model into an MPT-RT model, an additional probability parameter L_j is simply added to each branch of the MPT model such that branch j terminates in a fast response with probability L_j and in a slow response with the complementary probability $(1 - L_j)$[1]. Importantly, MPT-RT models allow to include equality constraints and order restrictions between latency parameters to test different hypotheses on the relative speed of processing branches.

In the following, we present the four process models that Klauer and Voss (2008) discussed for the WIT in more detail and show how they can be formalized as MPT-RT models. To foreshadow, the four process models are formalized as specific versions of the PDP that rely on different parametrizations (with A conditional on C for PCRM and GM, and C conditional on A for DIM and PCM) and they differ in their latency assumptions for the processing branches. The corresponding MPT-RT model equations are listed in the online supplementary material. Table 1 presents an overview of the process models and their key assumptions. In what follows, we use the uppercase letter

[1] The latency parameter L_j may vary as a function of prime race, target object presented (i.e., $G =$ gun; $T =$ tool) and the previous process parameter in the branch the latency parameter belongs to (i.e., C, A, $(1 - C)$, or $(1 - A)$). Subscripts + and - indicate that the parameter is conditional on A or $(1 - A)$, respectively.

M with a subscript indicating the specific model to denote the proposed MPT-RT implementations of the respective process models (i.e., M_{PCRM}, M_{DIM}, M_{PCM}, and M_{GM}).

Insert Table 1 about here

Specifically, according to the *preemptive conflict-resolution model (M_{PCRM})* (Figure 2), an initial binary decision determines whether the response to a stimulus is driven by controlled processes with probability C or by automatic processes with probability $(1 - C)$. Importantly, the M_{PCRM} assumes that participants adjust their reliance on the controlled versus the automatic process preemptively on a trial-by-trial basis. For example, participants might try to respond more carefully after committing an error by relying more on the controlled process than the automatic process in the upcoming trial. Conversely, participants might choose to rely more often on automatic processes if cognitive resources are scarce or if they are under time pressure (Klauer & Voss, 2008). To reiterate, the controlled process necessarily results in a correct response, whereas the automatic process corresponds to the automatically activated associations of the prime that trigger the response *gun* with probability A and the response *tool* with the complementary probability $(1 - A)$ for both target conditions.

Insert Figure 2 about here

Regarding processing times, the M_{PCRM} assumes that automatic processes are equally fast, irrespective of the activated associations. Controlled processes are effortful and thus need more time to determine the response than spontaneous, prime-triggered automatic processes. When formalized in the MPT-RT framework, both automatic processing branches are assumed to be equally fast within

each prime-target condition ($L_A = L_{(1-A)}$). Furthermore, latency parameters of the automatic processing branches (L_A and $L_{(1-A)}$) are expected to be larger compared to the latency parameter of the controlled processing branch (L_C).

In contrast to the M_{PCRM}, the *default interventionist model (M_{DIM})* lacks a preemptive conflict-resolution process. According to the M_{DIM} (Figure 3), the automatic process suggests, based on activated associations, a default response towards *gun* with probability A or towards *tool* with probability $(1 - A)$. This default response may then be overcome by the controlled process with probability C, resulting in a correct response. In contrast, if the default response is not overcome by the controlled process with probability $(1 - C)$, the default response suggested by the automatic process is given. The DIM presupposes that the overcoming probabilities of controlled processes do not depend on whether the automatically suggested default response is *gun* (overcoming probability $C+$) or *tool* (overcoming probability $C-$). Hence, there is only a single process parameter for each prime condition $C = C+ = C-$ that characterizes controlled processes in the M_{DIM}.

Insert Figure 3 about here

Regarding processing times, the M_{DIM} assumes that overcoming the default response is effortful and thus takes more time. In contrast, when the controlled process fails to intervene, the response follows a fast, spontaneous, prime-driven automatic process. In the MPT-RT framework, both processing branches that lead to the default response are assumed to be equally fast within each prime-target condition ($L_{(1-C+)} = L_{(1-C-)}$), and the same holds for both controlled processing branches ($L_{(C+)} = L_{(C-)}$). Furthermore, latency parameters of the default response branches ($L_{(1-C+)}$ and $L_{(1-C-)}$) are expected to be larger than those corresponding to the controlled processing branches ($L_{(C+)}$ and $L_{(C-)}$).

The *parallel competitive model (M_{PCM})* (Figure 4) resembles the M_{DIM} in that the automatic process suggests, based on activated associations, a default response towards *gun* with probability A or towards *tool* with probability $(1 - A)$. Other than in the previously discussed models, however, the controlled process runs in parallel to the automatic process in the M_{PCM}. When either of the two parallel processes terminates, it suggests a response. If the automatic process proposes a response congruent with the target condition (with probability A for the gun target and $(1 - A)$ for the tool target condition), it leads to a correct response in congruence with the controlled process. If the automatic process proposes a response incongruent with the target condition (with probability $(1 - A)$ for the gun target and A for the tool target condition), the conflict is resolved in favor of the controlled process with probability C and in favor of the automatic process with probability $(1 - C)$.

Insert Figure 4 about here

Regarding processing times, the M_{PCM} assumes that processing branches characterized by congruency of suggested responses are fast (i.e., $L_{G,A}$ and $L_{T,(1-A)}$) whereas processing branches characterized by conflict between processes are equally slow (i.e., $L_{G,C} = L_{G,(1-C)}$ and $L_{T,C} = L_{T,(1-C)}$).

Finally, according to the *guessing model (M_{GM})*, the response to a stimulus is either driven by a controlled or an automatic process with probability C and $(1 - C)$, respectively. The basic processing-tree structure of the M_{GM} is equivalent to the M_{PCRM} (Figure 2) and thus does not require a new figure for purposes of illustration. However, both the psychological interpretation and the restrictions imposed on the parameters differ between the M_{GM} and the M_{PCRM}, making them distinguishable in terms of their empirical predictions. In the M_{GM}, the controlled process succeeds in determining the correct response with probability C. If the controlled process fails with probability $(1 - C)$, the response is guessed as *gun* or *tool* with probabilities A and $(1 - A)$, respectively, depending

on stereotypes activated by the prime. Regarding processing times, the M_{GM} assumes that the controlled process results in faster responses compared to those of the automatic processing branches because additional guessing is involved in the latter. In the MPT-RT framework, both automatic processing branches are assumed to be equally fast within each prime-target condition ($L_A = L_{(1-A)}$) and latency parameters of the automatic processing branches (L_A and $L_{(1-A)}$) are expected to be smaller compared to the latency parameter of the controlled processing branch (L_C).

In sum, although the M_{GM} and the M_{PCRM} are equivalent in their MPT model equations, they differ with respect to the order restrictions imposed on the latency parameters. Specifically, automatic processing branches are faster in the M_{PCRM} ($L_C < L_A = L_{(1-A)}$) whereas controlled processing branches are faster in the M_{GM} ($L_C > L_A = L_{(1-A)}$).

Overview of Model Comparisons

As explained above, all four psychological process models are instantiations of the PDP but make different assumptions about the nature and the chronological order of the underlying cognitive processes. Hence, these models cannot be distinguished based on the accuracy of responses across the four prime-target conditions alone. However, because these process models impose different constraints on the latency parameters for each branch within the MPT-RT framework, they can be tested against each other, provided that both response frequencies and response latencies are available for all prime-target conditions of the WIT.

An appropriate framework for testing the four process models is response time-extended multinomial processing tree modeling (MPT-RT; Heck & Erdfelder, 2016). MPT-RT models enable the joint estimation of the core parameters C and A as well as the relative branch latencies L_j. Moreover, they allow researchers to test different sets of equality constraints and order restrictions on the parameters (Knapp & Batchelder, 2004) as those implied by the four process models. In contrast, previous work had to rely on comparisons of observed overall response times for correct and false responses across conditions (Klauer & Voss, 2008), that is, response times aggregated across different processing branches that terminate in the same response.

In what follows, we first evaluate the MPT-RT model fit of the psychological process models by re-analyzing eight previously published WIT data sets (Model Comparison 1). This allows us to identify whether there are specific process models that are in line with the available data originating from different samples, stimuli, and researchers applying the WIT. Second, we compare different sets of additional restrictions on the branch latencies and try to identify more parsimonious model versions that describe the observed data with fewer parameters (Model Comparison 2). Based on a simplifying restriction pattern that is in line with our data, we propose two extended MPT-RT models that discriminate between the best-fitting models of the previous comparison – the M_{PCRM} and the M_{DIM} – using plausible auxiliary assumptions (Model Comparison 3). For the sake of brevity, we focus on goodness-of-fit measures of the estimated models only. Group level parameter estimates, and their corresponding descriptive statistics are provided in the online supplementary material for Model Comparisons 1, 2, and 3 (https://osf.io/7vjrq).

Model Comparison 1: Testing the Fit of the Four Process Models

In Model Comparison 1, we estimated MPT-RT implementations of the four psychological process models M_{PCRM}, M_{DIM}, M_{PCM}, and M_{GM} (see Figures 2 to 4) with order restrictions on the parameters as listed in Table 1. The equations for each model can be found in the online supplementary materials. We performed model comparisons across eight data sets of the WIT as described in the next section.

Method

Study Search and Inclusion Criteria. Starting from the set of studies included in the meta-analysis by Rivers (2017), we selected North American studies that made use of the standard four-condition WIT paradigm, that is, investigated cross-classified White male faces and Black male faces as primes with guns and tools as target objects in 2 x 2 within-subject designs. We obtained a total of eight relevant data sets originating from five publications. The respective data sets were available either via the Open Science Framework (i.e., Madurski & LeBel, 2015; Rivers, 2017), or obtained by contacting the first or second author (i.e., Amon & Holden, 2016; Correll, 2008; Lambert et al., 2003). Table 2 provides an overview of the main characteristics of each data set such as the

number of participants (N_{sample} = 22 – 50), the percentage of Caucasian participants (17 % – 83 %), the number of trials (N_{trials} = 200 – 1,100), prime duration (200 ms – 300 ms), and the response time limit (500 ms – 1,000 ms). All data sets were obtained in the USA or in Canada and include responses from primarily non-black university students as participants. We used the data sets in the same way as they were used in the original publications and provided by the authors; no further participants were excluded. The single exception was the study by Lambert et al. (2003) for which only a part of the data set included response times in addition to response frequencies. If a data set contained additional experimental manipulations (Amon & Holden, 2016; Correll, 2008; Lambert et al., 2003; Madurski & LeBel, 2015) or additional prime stimulus materials (Rivers, 2017) compared to the standard procedure of the WIT (Payne, 2001), we excluded these trials from data analysis.

Insert Table 2 about here

MPT-RT Modeling Procedure. We used the hierarchical latent-trait model of Klauer (2010) as the statistical framework for all analyses. This model family allows for variability of all parameters across individuals such that the probit-transformed person parameters follow a joint multivariate normal distribution at the group level. In contrast to alternative hierarchical MPT models, the latent-trait model not only provides both individual and group-level parameters (i.e., parameter means) but also variances and correlations of person parameters across participants.

For assessing goodness of fit, we used the Bayesian posterior predictive p-value based on the test statistics *T1* and *T2* (Klauer et al., 2015; Klauer, 2010). *T1* summarizes how well the model accounts for the average response frequencies across participants. It resembles the Pearson goodness-of-fit statistic χ^2 often used in categorical data analysis (Klauer et al., 2015; Riefer & Batchelder, 1991). The test statistic *T2*, in contrast, summarizes how well the model accounts for the variances and correlations of the response frequencies across participants. The posterior predictive p-value

represents the comparison of the calculated test statistics obtained for observed and predicted response frequencies. A value of $p > .05$ is typically seen as evidence that model assumptions are in line with the data (Klauer et al., 2015; Klauer, 2010).

We used the R package TreeBUGS (Heck et al., 2018) to fit hierarchical versions of the four MPT-RT models to the eight data sets of interest. The Markov Chain Monte Carlo (MCMC) algorithm was run for three independent estimation chains with 750,000 iterations each, of which 250,000 were removed as a burn-in period. Every 50th iteration was retained to compute summary statistics. For models M_{PCRM}, M_{DIM}, M_{PCM}, and M_{GM}, the parameters A and C were allowed to vary between primes (i.e., Black vs. White male face primes), and latency parameters L_j were allowed to vary between primes, target object conditions (i.e., guns vs. tools), and the previous process parameter in the branch the latency parameter belongs to (i.e., C, A, $(1 - C)$, or $(1 - A)$). The Rubin-Gelman statistic $\hat{R}$ was smaller than 1.05 for all parameter estimates across all models, showing an acceptable convergence of MCMC sampling.

Response-Time Data Preparation. Regarding response times, we included all trials provided by the authors of the data sets, that is, all trials with response times faster than the response-time limit of the corresponding data set (between 500 ms and 1000 ms). To estimate MPT-RT models, we divided trials into fast and slow responses for each participant based on whether a specific response was faster or slower than the individual mean of their log-transformed response times across response categories (see Heck & Erdfelder, 2016). This gives us two discrete RT bins of fast and slow responses for every correct and false response across conditions, separately for each individual. Although categorization of response times results in a loss of information compared to analyzing continuous response times, participant-specific categorization in fast and slow responses is advantageous because estimation is simpler, more robust, and does not need any a priori assumptions about the distribution of response times (Heck & Erdfelder, 2016).

Insert Table 3 about here

Results

Table 3 shows the goodness-of-fit measures for the four psychological process models M_{PCRM}, M_{DIM}, M_{PCM}, and M_{GM} including the respective order restrictions on the latency parameters for each data set. Regarding model fit statistics, models M_{PCRM} and M_{DIM} show acceptable fit in both test statistics for almost all data sets (*p(T1)*s $> .218$; *p(T2)*s $> .177$). The single exception is Study 1a of Amon and Holden (2016) with misfit in the *T2* statistic (*p(T1)* $< .005$). In stark contrast, models M_{PCM}, and M_{GM} show blatant misfit in both test statistics for all data sets (all *p(T1)*s $< .001$; all *p(T2)*s $< .008$).

Discussion

The data clearly favor the preemptive conflict-resolution model M_{PCRM} and the default interventionist model M_{DIM}. In contrast, the parallel competitive model M_{PCM} and the guessing model M_{GM} are clearly rejected by the data. The reason for the good fit of M_{PCRM} and M_{DIM} is the typically observed data pattern that WIT error latencies are faster than correct response latencies. While this pattern can be accommodated by M_{PCRM} and $M_{DIM,}$ it conflicts with M_{PCM} and M_{GM}. Hence, the joint modeling of response frequencies and response times results in the same substantive conclusions as the model-free analysis of response times by Klauer and Voss (2008), namely, that automatic and controlled processing routes are fast and slow, respectively, in line with the PCRM and the DIM. Specifically, automatic processes lead to both fast incorrect and fast correct responses as a consequence of stereotypical associations induced by the prime's race. In contrast, controlled process branches involve more time and effort to identify the target object correctly (i.e., target discrimination for the PCRM and target discrimination with default response inhibition for the DIM).

Importantly, despite different theoretical assumptions underlying each model, M_{PCRM} and M_{DIM} are mathematically equivalent if independence of automatic and controlled processes is assumed,

that is, if controlled process parameters (and their corresponding branch latency parameters) do not differ between congruent and incongruent default responses in the M_{DIM} (i.e., $C = C+ = C-$, see Figure 3) as is the case here. Hence, the latencies of the $C+$ and $C-$ branches cannot be estimated separately. In other words, to distinguish the influence of congruent versus incongruent automatic responses on branch latencies in $M_{DIM,}$ additional assumptions about the latency parameters need to be incorporated.

To gain more degrees of freedom for tests between the M_{PCRM} and the M_{DIM}, we opted to search for a more parsimonious model (Model Comparison 2). Using this approach, influences of congruent versus incongruent automatic responses on branch latencies can be estimated separately (Model Comparison 3) which in turn allows for a test between M_{PCRM} and M_{DIM}.

Model Comparison 2: Constraining Process Latencies Across Conditions

The M_{PCRM} and the M_{DIM} previously analyzed in Model Comparison 1 provide for separate controlled and automatic branch latency parameters for each prime-target condition. Here, we consider the possibility that this assumption is unnecessarily complex because corresponding branch latencies may not differ between prime or target conditions or even both.

For automatic process routes, we assume that latencies parameters do not differ between prime-target conditions. This is reasonable because the automatic associations are instantly triggered by the prime and thus do not depend on the target object that follows later. Moreover, even if relative response latencies of automatic process branches should differ slightly between prime conditions, we expect these differences to be negligible compared to the relatively slow controlled process branch latencies.

For the controlled process routes, we assume that response latencies may differ between gun and tool targets while being independent of prime condition (i.e., whether a Black or White face was displayed before). This assumption is plausible for two reasons. First, we expect the speed of controlled processes to vary between target object conditions because previous WIT research has shown that participants' correct responses tend to be faster for gun than for tool targets (Payne, 2001;

Rivers, 2017). Second, equality of latency parameters within target condition is in line with the invariance assumption of the PDP for the controlled process parameters. This invariance assumption presupposes that target discrimination is not affected by prime race (Payne et al., 2005). For a full discussion of the validity and possible criticism of the invariance assumption, see the General Discussion section.

To test this restriction pattern in the MPT-RT framework, we propose model M_{Basic} as a starting point. Model M_{Basic} includes the mathematically equivalent models M_{PCRM} and M_{DIM} as submodels (see online supplementary material). More precisely, the MPT-RT model equations of M_{Basic} are equivalent to those submodels but do not impose any order restrictions on the latency parameters (as those listed in Table 1). Starting from M_{Basic}, the more parsimonious nested model M_{Pars} imposes additional equality restrictions on automatic and controlled processing branches as outlined above. Specifically, for the automatic processing branches, latency parameters are equated across all four prime-target conditions. This refers to parameters L_A and $L_{(1-A)}$ for M_{PCRM} (see Figure 2) and latency parameters $L_{(1-C+)}$ and $L_{(1-C-)}$ for M_{DIM} (see Figure 3), respectively. For the controlled processing branches, latency parameters are equated within target conditions. This refers to parameter L_C for M_{PCRM} and latency parameters $L_{(C+)}$ and $L_{(C-)}$ for M_{DIM}, respectively.

Method

We again fitted the hierarchical MPT models using TreeBUGS (Heck et al., 2018). For comparisons between models taking model complexity into account, we used the Deviance Information Criterion (DIC; Klauer et al., 2015; Spiegelhalter et al., 2002) and the Widely Applicable Information Criterion (WAIC; Vehtari et al., 2017; Watanabe, 2010). The DIC is a hierarchical modeling generalization of the Akaike Information Criterion (AIC). The DIC consists of a term quantifying lack of model fit and a term penalizing model complexity (Klauer et al., 2015; Klauer, 2010; Spiegelhalter et al., 2002, 2014). The WAIC can be seen as an improved version of the DIC, as it is fully Bayesian and based on the entire posterior distribution of participants and not solely on a point estimate. The WAIC consists of a term summed across the estimated log pointwise predictive density and a penalizing term summed across the estimated variance of the log pointwise predictive density (Vehtari

et al., 2017). Like AIC, the model with the lowest DIC or WAIC value, respectively, represents the best compromise between fit and complexity, with an absolute ΔDIC (ΔWAIC) difference larger than two usually interpreted as evidence for one model compared to the other (Burnham & Anderson, 2004; Klauer et al., 2015; Spiegelhalter et al., 2002).

We used the same estimation settings as in Model Comparison 1. The Rubin-Gelman statistic $\hat{R}$ was smaller than 1.05 for all parameter estimates across all models. To obtain stable goodness-of-fit measure statistics, we estimated each model ten times and calculated means and standard errors for the model fit statistics *p(T1)*, *p(T2)*, DIC, and WAIC.

Results

Table 4 shows the model fit statistics *p(T1)* and *p(T2)* of M_{Basic} and M_{Pars}, as well as DIC and WAIC differences between models. Model M_{Basic} fits all data sets (all *p(T1)*s $>$ .473; all *p(T2)*s $>$ .163). The parsimonious model M_{Pars} fits almost all data sets (all *p(T1)*s $>$ .144; all *p(T2)*s $>$ .084) except Study 1a of Amon and Holden (2016) and Study 1b of Rivers (2017) (*p(T1)*s $<$ .037).

ΔDIC (ΔWAIC) was calculated as the difference of the DIC (WAIC) measures between M_{Pars} and M_{Basic}. Hence, a negative ΔDIC (ΔWAIC) supports M_{Pars} relative to M_{Basic}. As evident from Table 4, for five data sets DIC favors M_{Pars} more than M_{Basic} (ΔDICs $<$ -5.35). For three data sets no model is favored (ΔDICs $<$ |1.52|). With respect to WAIC, two data sets favor M_{Pars} more than M_{Basic} (ΔWAICs $<$ -9.77) while preferences are reversed for five data sets (ΔWAICs $>$ 17.05) and for one data set no model is favored (ΔWAIC $<$ |0.29|).

In sum, both M_{Basic} and M_{Pars} fit almost all data sets. While model comparisons via DIC revealed a preference for M_{Pars}, those based on WAIC showed the reversed pattern.

Insert Table 4 about here

Discussion

Overall, the parsimonious model M_{Pars} fits the data almost as well as the more general model M_{Basic}. Although model comparison via DIC and WAIC diverged slightly, the posterior predive *p*-values for *T1* and *T2* showed that the nested model M_{Pars} does a good job in describing both the observed frequencies and the covariance structure of the responses for nearly all data sets[2]. For the weapon identification task, this implies that responses based on automatic processes tend to be equally fast irrespective of elicited automatic response and prime race. However, responses based on the controlled process may depend on the to-be-identified target. This may be due to differences between stimuli in terms of visual discriminability. Alternatively, it may also reflect participants' strategy to respond to a threatening weapon target as soon as they identified it but be more hesitant when they believed to identify a tool, primarily to avoid missing a weapon target. This is in line with previous findings showing that participants tend to respond faster to threat-related targets (Payne, 2001; Rivers 2017).

For follow-up model comparisons between M_{PCRM} and M_{DIM} (Model Comparison 3), we rely on the latency-parameter restrictions across conditions imposed by M_{Pars} as auxiliary assumptions. The additional auxiliary assumptions provide a more parsimonious model with additional degrees of freedom. This allows us to test between different order restrictions for the controlled processing branches implied by the PCRM and the DIM.

Model Comparison 3: Testing Extended Versions of the PCRM and the DIM

Model M_{PCRM} and M_{DIM} are mathematically equivalent if controlled process parameters (and their corresponding branch latency parameters) do not differ between congruent and incongruent default responses in model M_{DIM}, that is, if $C = C+ = C-$ (Figure 3). To distinguish between these

[2] Additional model comparisons based on nested latency parameter restrictions for the automatic process routes (across prime conditions and prime-target conditions) and for the controlled process routes (across target conditions and prime-target conditions) are reported in the online supplementary materials.

two models, further assumptions about the latency parameters are necessary. According to Klauer and Voss (2008), the DIM is likely to be more appropriate for the WIT than the PCRM because the DIM can explain faster identification of weapons after Black face primes, even when the success probability of controlled processes is high. This data pattern can be explained by assuming that controlled processes are faster following default responses congruent with the target condition compared to incongruent default responses. As outlined by Klauer and Voss (2008), the PCRM lacks a plausible mechanism to explain this outcome.

To test this idea, we combined the order constraints of the two well-fitting models of Model Comparison 1 (M_{PCRM} and M_{DIM}) with the equality constraints of the parsimonious model M_{Pars} from Model Comparison 2. To reiterate, according to M_{Pars} the relative speed of automatic processing branches does not differ between different prime-target conditions, and the relative speed of controlled process branches does not differ within target conditions. By combining the different sets of order constraints of M_{PCRM} and M_{DIM} (Table 1) with the equality constraints of M_{Pars}, two extended model versions $M_{PCRM,ext}$ and $M_{DIM,ext}$ are obtained. In contrast to the model version considered in Model Comparison 1, $M_{PCRM,ext}$ and $M_{DIM,ext}$ have less parameters than degrees of freedom available in the data. Goodness-of-fit differences between both models can therefore be assessed.

Model $M_{PCRM,ext}$ mirrors the MPT-RT structure of M_{PCRM} model (Figure 2) but imposes the additional constraints that automatic processing branch latencies do not differ between prime-target conditions and that controlled processing branch latencies do not differ within target conditions. Regarding order restrictions for latency parameters, $M_{PCRM,ext}$ assumes that automatic processing branches are generally faster than controlled processing branches in each target condition (i.e., $L_A > L_{G,C}$ and $L_A > L_{T,C}$).

Analogously, model $M_{DIM,ext}$ has the same MPT-RT structure as M_{DIM} (Figure 3) but imposes the additional constraints that latencies of automatic processing branches do not differ between prime-target conditions and that latencies of controlled processing branches do not differ within target conditions. Regarding order restrictions for latency branches, recall that model M_{DIM} assumes that controlled processing branches are equally fast ($L_{C+} = L_{C-}$), regardless of whether they follow a default

response that is congruent or incongruent with the target condition. In contrast, the extended model $M_{DIM,ext}$ assumes that controlled processes following congruent default responses are faster compared to those following incongruent default responses. In sum, according to $M_{DIM,ext}$, automatic processing branches are generally faster than controlled processing branches for congruent responses in each target condition. These, in turn, are faster than controlled processing branches for incongruent responses in each target condition (i.e., $L_A > L_{G,C+} > L_{G,C-}$ and $L_A > L_{T,C-} > L_{T,C-}$).

Method

We used the same methods and estimation settings as in Model Comparison 1 and 2. $\hat{R}$ was smaller than 1.05 for all parameter estimates across all models. Again, to obtain stable goodness-of-fit statistics, we estimated each model ten times and calculated means and standard errors for the model fit statistics *p(T1)*, *p(T2)*, DIC, and WAIC.

Results

Table 5 shows the model fit statistics *p(T1)* and *p(T2)* of $M_{PCRM,ext}$ and $M_{DIM,ext}$, as well as DIC and WAIC differences between models. Model $M_{PCRM,ext}$ fits almost all data sets (all *p(T1)s* > .146; all *p(T2)s* > .082), except for Study 1a of Amon and Holden (2016; *p(T1)* = .004 and *p(T2)* = .031) and Study 1b of Rivers (2017; *p(T1)* < .001). Similarly, model $M_{DIM,ext}$ fits almost all data sets (all *p(T1)*s > .209 and all *p(T1)*s > .157), except for Study 1a of Amon and Holden (2016; *p(T1)* = .004 and *p(T2)* = .031).

The ΔDIC (ΔWAIC) was calculated as the difference of the DIC (WAIC) measures between $M_{DIM,ext}$ and $M_{PCRM,ext}$. Hence, a negative ΔDIC (ΔWAIC) supports $M_{DIM,ext}$ relative to $M_{PCRM,ext}$. Note that we chose to not interpret ΔDIC and ΔWAIC for Study 1a of Amon and Holden (2016) as neither $M_{PCRM,ext}$ nor $M_{DIM,ext}$ fits the data well. For the remaining seven data sets, DIC favors $M_{DIM,ext}$ more than $M_{PCRM,ext}$ in five cases (ΔDICs < -3.38) and WAIC even in six cases (ΔWAICs < -3.20). While DIC preferences were reversed for one data set (ΔDICs = 2.71), we observed no preference reversals for WAIC. None of the models is favored for one data set, both with respect to DIC (ΔDICs = -1.35) and WAIC (ΔWAICs = -0.14).

Insert Table 5 about here

Discussion

In sum, both models $M_{PCRM,ext}$ and $M_{DIM,ext}$ fit almost all data sets. This notwithstanding, there is a general preference for $M_{DIM,ext}$, irrespective of the model selection measure used. The only exceptions are studies by Amon and Holden (2016) which show either no fit (Study 1a) or no preference for $M_{DIM,ext}$ (Study 1b). These discrepancies are perhaps due to a significantly larger number of trials (N_{trials} = 1,100) in Amon and Holden's studies compared to the other studies (all $N_{trials} \leq 384$). The total number of trials may affect the general strategy how participants perform in the WIT.

The preference for $M_{DIM,ext}$ indicates that automatically activated default responses always interfere with weapon identification in every trial. According to $M_{PCRM,ext}$, in contrast, only some trials are influenced by automatic processes. Regarding controlled responding, $M_{DIM,ext}$ maintains that two processes play an important role in weapon identification, namely, (1) target discrimination and (2) overcoming the default response when the latter turns out to be incongruent with the to-be-identified target.

General Discussion

The *weapon identification task* (WIT) is a widely used sequential priming procedure to assess the influence of racial threat stereotypes on weapon identification (Payne, 2001; Rivers, 2017). The *process dissociation procedure* (PDP) has routinely been used as measurement model for the WIT to estimate the influence of controlled and automatic processes on task performance and to assess how these processes are affected by different factors. For example, participants show less controlled processing for shorter response-time windows (Payne, 2001; Payne et al., 2002), when more anxious because they expect public accountability for the own performance (Lambert et al., 2002), and when their cognitive resources are limited (Govorun & Payne, 2006; Payne, 2005; Ito et al., 2015).

Furthermore, the automatic association of Black male face primes with guns is enhanced when race is made salient (Payne et al., 2002) and participants are in a positive mood (Huntsinger et al., 2009) or can be diminished when participants use implementation intentions to think safe for Black face primes (Stewart & Payne, 2008).

Although the proportion of responses determined by controlled or automatic processes might be influenced by these factors, their nature, relationship, and temporal order remains underdetermined in the standard PDP framework. For example, the controlled process can entail target discrimination and overcoming of automatic associations while the automatic process can entail activation of racial associations or a prime-informed guessing tendency. In fact, the PDP is consistent with four process models for the WIT: the preemptive conflict-resolution model (PCRM), the default interventionist model (DIM), the guessing model (GM), and the parallel competitive model (PCM).

In this article, we formalized these process models in the framework of response time-extended multinomial processing tree (MPT-RT) models (Heck & Erdfelder, 2016). This approach has several advantages. First, MPT-RT models combine information on response frequencies and response times in the same model. Second, MPT-RT models provide model fit and model selection statistics to test between (or select among) models. Third, MPT-RT models provide for equality and order restrictions on both the structural and the latency parameters of the respective MPT-RT model. Most importantly in the present context, the MPT-RT framework allows to test the four process models of the WIT against each other. This is not possible when analyzing response frequencies in isolation (as in the standard PDP). Essentially, the different process models become empirically distinguishable after extending the PDP model with latency parameters for fast and slow responses in different processing branches. The major advantage is that model fit can be assessed for response frequencies and response latencies conjointly rather than focusing on each dependent variable in two separate steps.

Testing between Psychological Process Models

Our results support the default interventionist model (DIM) and the preemptive conflict-resolution model (PCRM), but not the guessing model (GM) and the parallel competitive model

(PCM) for the WIT. For all eight data sets considered here, the DIM and PCRM fitted the data very well. Based on the parameter estimates we observed, both models predict controlled process branches to be slower than automatic process branches. This is in line with the result pattern typically observed for the WIT (i.e., fast errors and slow correct responses).

Further model comparisons based on additional invariance assumptions deemed the DIM as a more appropriate process model for the WIT than the PCRM. This is in line with the reasoning of Klauer and Voss (2008) who argued that the DIM can explain that, in studies with high accuracy rates, response bias due to prime race persists in response latencies. According to the DIM, overcoming the default response by controlled processing leads to slower responses if automatic and controlled processes disagree than if both agree. The PCRM cannot explain this pattern easily because the latency of the controlled process is assumed to be independent of the outcome of the automatic process.

Additional support for the DIM comes from research on brain region activation after seeing Black and White faces (Cunningham et al., 2004). Specifically, short presentation (30 ms) of faces leads to higher activation of the amygdala for Black compared to White faces, a brain region typically responsive to the emotionality of a stimuli (e.g., induced threat). However, a longer presentation time (525 ms) of Black and White faces leads to no difference in amygdala activation. This lack of difference in amygdala activation for longer presentation times of faces is associated with increased activity of the dorsolateral prefrontal cortex and the anterior cingulate, brain regions associated with controlled processing and executive function. In line with the DIM, this suggests that, in a first step, automatic associations and emotional reactions are instantly activated when processing a face stimulus while in a second step (about half a second later) these associations can be inhibited and modulated by controlled processes.

Adopting the DIM as the best-fitting model of cognitive processes underlying performance in the weapon identification task thus implies that automatic threat-stereotype associations interfere with object identification throughout each trial. Hence, controlled processes are required for both object discrimination *and* conflict resolution when the automatic process elicits an incongruent default

response. In other words, object discrimination abilities and cognitive control for overcoming default responses *both* play a crucial role in WIT performance.

Psychological Implications

The assumption of automatically generated default responses in the DIM nicely fits findings regarding the malleability of automatic racial stereotyping depending on participants' latent focus on race at the beginning of each trial. For example, race salience (Payne et al., 2002; Jones & Fazio, 2010) strengthens the association of Black males with guns, whereas focusing on other face characteristics (e.g., age) diminishes the association of Black males with guns (Jones & Fazio, 2010; Todd et al. 2021). Similarly, implementation intentions (i.e., previously determined if-then action plans) allow to adjust associations (e.g., as safe) for specific face primes (Stewart & Payne, 2008). Hence, even if face primes elicit a default response at the beginning of each trial, this response is malleable if the associations tied to the face primes are adjusted in advance.

Similarly, findings regarding reduced capabilities of participants for correct responding are easily explained by failures of the controlled process in target identification and overcoming bias. For example, shorter response deadlines, stress, and diminished cognitive capabilities are known to compromise target discrimination as well as conflict resolution to overcome bias (Govorun & Payne, 2006; Lambert et al., 2002; Payne, 2001; Payne et al., 2002; Payne, 2005; Ito et al., 2015). Furthermore, this is in line with the findings that participants with higher cognitive inhibitory capabilities (Ito et al., 2015) or with higher internal motivation to control their prejudice (Volpert-Esmond et al., 2020) show more controlled responding[3]. This suggests that more practice in resolving conflicts induced by automatic default responses results in more correct responding. Hence, strategy

[3] As pointed out by a reviewer, these results also align with the PCRM. From the perspective of the PCRM, individuals with better inhibitory capacities or higher prejudice-control motivation may simply be the ones who choose more often to preempt conflict by engaging in controlled processing.

training and interventions that strengthen object discrimination *and* cognitive conflict resolution should improve overall performance in the WIT.

MPT-RT implementations of WIT process models such as the DIM provide a number of advantages that can guide future research. For example, participants of different social groups or participants working under different experimental conditions can be compared not only with respect to their cognitive process parameters but also with respect to the relative speed of automatic and controlled processing branches. Moreover, given the hierarchical nature of the MPT-RT approach proposed here, personality traits or cognitive abilities can be studied as external covariates and potential predictors of cognitive process and latency parameters. Both of these approaches are helpful not only for basic social cognition research but also for designing and evaluating training programs and interventions that counteract maladaptive effects of social stereotypes.

Limitations

One limitation when using MPT-RT models of WIT performance is the necessity to constrain model parameters to achieve model identifiability. Regarding latency parameters, models M_{PCRM}, M_{DIM}, and M_{GM} assume that relative latencies do not differ between the two automatic processing branches suggesting a *gun* or a *tool* response. Similarly, according to model M_{PCM} relative latencies do not differ between the two processing branches that represent conflicts between incongruent default responses and successful target discrimination. Since the models M_{PCRM}, M_{DIM}, M_{PCM}, and M_{GM} already have as many free parameters as there are independent response categories in the data (i.e., $df = 0$), they do not allow to test these underlying assumptions.

In a similar vein, the comparison of models $M_{PCRM,ext}$ and $M_{DIM,ext}$ is based on auxiliary assumptions for latency parameter restrictions across conditions. The parsimonious model M_{Pars} showed acceptable model fit, but model selection indices DIC and WAIC do not unequivocally support M_{Pars} relative to the unrestricted model M_{Basic}. Hence, it should not be overlooked that the model comparison of the extended models $M_{PCRM,ext}$ and $M_{DIM,ext}$ is conditional on the auxiliary assumptions implemented in M_{Pars}.

Furthermore, $M_{DIM,ext}$ in Model Comparison 3 is endowed with the capacity to account for stereotype-congruency effects of correct response times via latency parameters whereas $M_{PCRM,ext}$ cannot. More specifically, $M_{DIM,ext}$ differentiates between latencies of controlled process branches resulting in congruent versus incongruent default responses whereas the $M_{PCRM,ext}$ does not have this feature. Alternatively, the stereotype-congruency effect in response times might be due to a variation in latencies of automatic process branches, meaning that the speed of stereotype-congruent automatic process branches is faster than the speed of incongruent automatic process branches. For example, the association of a Black face with a gun occurs faster than the association of a Black face with a tool (Klauer & Voss, 2008). This alternative view allows to explain stereotype-congruency effects based on the latency parameters of the automatic process branches. It can also account for the reversed congruency effect reported by Klauer and Voss (2008), that is, that errors for tool targets are faster if they followed Black face primes and errors for gun targets are faster if they followed White face primes.

Model comparisons implementing this alternative specification for M_{PCRM} and M_{DIM} (i.e., stereotype-congruent automatic process branches are faster than incongruent ones) are listed in the online supplemental materials as Model Comparison 4 and 5 (https://osf.io/7vjrq). Importantly, these alternative model comparisons overall result in the same conclusion as reported above, namely, that the DIM is more closely aligned with the available data than the PCRM. Hence, the specific assumptions implemented in our Model Comparison 3 are not crucial for the conclusion that the DIM provides the best account of WIT response frequencies and response times when considered jointly.

Beyond the constraints on the latency parameters, the PDP itself has been criticized because of the assumption that the controlled process parameter C is invariant across target conditions. Payne et al. (2005) as well as Klauer and Voss (2008) critically discussed this invariance assumption. For example, a metal tube as part of a target object might more easily be interpreted as part of a gun after seeing a Black face prime and more easily as part of a screwdriver after seeing a White face prime. Consequently, the prime would be expected to modulate the controlled process of discriminating the object (Klauer et al., 2015; Klauer & Voss, 2008; Payne et al., 2005). Experimental

tests of the invariance assumption resulted in mixed outcomes so far (Klauer et al., 2015; Payne et al., 2005). Based on an extended version of the PDP, Klauer et al. (2015) observed results inconsistent with the invariance assumption. More recently, however, Todd et al. (2021) analyzed the WIT using diffusion modeling as an alternative analytic approach to the PDP. Their diffusion analysis suggests that the discrimination process (as mirrored in the drift rate) was not influenced by an interaction between primes and targets, a result that is in line with the invariance assumption. Clearly, further research addressing the validity of the invariance assumption is needed.

Alternative modeling approaches

As alternatives to MPT-RT modeling, other cognitive modeling approaches can be used to gain a deeper understanding of underlying processes. As already mentioned, Todd et al. (2021) used the diffusion model (DM), a specific version of evidence-accumulation models, to analyze WIT performance. The DM decomposes decisions into four components: an initial response bias towards guns or tools, the quality of information extracted from a stimulus (named drift rate), non-decision time (e.g., encoding and motor response time), and the amount of evidence required to make a decision (named threshold separation). In their DM analysis, they found that the initial response bias for Black face primes was closer to the gun response than for White face primes. The drift rate, however, did not vary as a function of prime race. This suggests that the prime biases the decision in the WIT from the outset of each trial, again a result that is in accordance with the DIM.

As an alternative to MPT-RT and DM approaches, RT-MPT models (Klauer & Kellen, 2018) could also be used to incorporate response times in multinomial models. RT-MPT models, in contrast to MPT-RT models, allow estimation of single-process latencies instead of relative latencies of entire process branches. To our knowledge, RT-MPTs have not been applied to the WIT so far. They could be a valuable alternative to gain more fine-grained insights into the processes underlying WIT performance. To conclude, future research on the WIT can benefit from using multiple modeling approaches to gain a deeper understanding of the underlying cognitive processes, to combine the strengths of the different approaches, and eventually achieve converging evidence across modeling approaches.

Conclusion

We reanalyzed eight data sets of the weapon identification task (WIT) to evaluate four process models of racial bias in weapon identification after formalizing them as response time-extended multinomial processing tree models. Our process model comparison clearly supported the default interventionist model (DIM) and the preemptive conflict-resolution model (PCRM). Both models share the assumption that responses based on automatic processes are faster than those based on controlled target identification. Follow-up model comparisons revealed a preference for the DIM vis-à-vis the PCRM. According to the DIM, automatic associations elicit response bias from the outset of each trial. Therefore, controlled processes – beyond their importance for target identification – are required to resolve possible response conflicts whenever default responses deviate from the correct response.

In addition to their significance for basic social cognition research, our results are highly relevant also for applied fields such designing and implementing police training. Our results suggest that interventions that increase both, the capacity for controlled processing to discriminate the target object *and* the ability to overcome default automatic responses, increase accuracy of weapon identification. In line with this conclusion, previous studies showed that increasing the capacity for controlled processing, for example, by providing sufficiently large response-time windows, by increasing cognitive resources, and by training to resolve conflicts induced by default responses (Govorun & Payne, 2006; Klauer & Voss, 2008; Kleiman et al., 2013; Lambert et al., 2002; Payne, 2001; Payne et al., 2002; Payne, 2005; Rivers, 2017; Ito et al., 2015), generally improves performance in the WIT. Under standard WIT conditions, however, people tend to be biased by race (or other social characteristics) early on in identifying innocuous or threatening objects. Nevertheless, by developing well-designed interventions, this bias likely can be overcome or at least reduced effectively. This is an important result, especially for bias awareness programs and police trainings that aim at reducing racial bias.

References

Amon, M. J., & Holden, J. G. (2016). *Fractal Scaling and Implicit Bias: A Conceptual Replication of Correll (2008)*. In A. Papafragou, D., Grodner, D., Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1553 – 1558). Austin, TX: Cognitive Science Society. https://cognitivesciencesociety.org/wp-content/uploads/2019/03/cogsci2016_proceedings.pdf

Bishara, A. J., & Payne, B. K. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology*, *45*(3), 524–534. https://doi.org/10.1016/j.jesp.2008.11.002

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Conrey, 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*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Correll, J. (2008). 1/f noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology*, *94*(1), 48–59. https://doi.org/10.1037/0022-3514.94.1.48

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological science*, *15*(12), 806-813. https://doi.org/10.1111/j.0956-7976.2004.00760.x

Evans, J. S. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339. https://doi.org/10.1080/13546780601008825

Govorun, O., & Payne, B. K. (2006). Ego—depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, *24*(2), 111-136.

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*(5), 1440–1465. https://doi.org/10.3758/s13423-016-1025-6

Huntsinger, J. R., Sinclair, S., & Clore, G. L. (2009). Affective regulation of implicitly measured stereotypes and attitudes: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *45*(3), 560–566. https://doi.org/10.1016/j.jesp.2009.01.007

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Jones, C. R., & Fazio, R. H. (2010). Person categorization and automatic racial stereotyping effects on weapon identification. *Personality and Social Psychology Bulletin*, *36*(8), 1073-1085. https://doi.org/10.1177/0146167210375817

Kahn, K. B., & Martin, K. D. (2020). The Social Psychology of Racially Biased Policing: Evidence-Based Policy Responses. *Policy Insights from the Behavioral and Brain Sciences*, *7*(2), 107–114. https://doi.org/10.1177/2372732220943639

Klauer, K. C. (2010). Hierarchical Multinomial Processing Tree Models: A Latent-Trait Approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221. https://doi.org/10.1037/xge0000044

Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology*, *82*, 111-130.

Klauer, K. C., & Voss, A. (2008). Effects of Race on Responses and Response Latencies in the Weapon Identification Task: A Test of Six Models. *Personality and Social Psychology Bulletin*, *34*(8), 1124–1140. https://doi.org/10.1177/0146167208318603

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*(4), 215–229. https://doi.org/10.1016/j.jmp.2004.03.002

Lambert, A. J., Payne, B. K., Jacoby, L. L., Shaffer, L. M., Chasteen, A. L., & Khan, S. R. (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology*, *84*(2), 277–295. https://doi.org/10.1037/0022-3514.84.2.277

Madurski, C., & LeBel, E. P. (2015). Making sense of the noise: Replication difficulties of Correll's (2008) modulation of 1/f noise in a racial bias task. *Psychonomic Bulletin & Review*, *22*(4), 1135–1141. https://doi.org/10.3758/s13423-014-0757-4

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192. https://doi.org/10.1037/0022-3514.81.2.181

Payne, B. K. (2005). Conceptualizing Control in Social Cognition: How Executive Functioning Modulates the Expression of Automatic Stereotyping. *Journal of Personality and Social Psychology*, *89*(4), 488–503. https://doi.org/10.1037/0022-3514.89.4.488

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology*, *38*(4), 384–396. https://doi.org/10.1016/S0022-1031(02)00006-9

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology*, *41*(1), 36–47. https://doi.org/10.1016/j.jesp.2004.05.001

Riefer, D. M., & Batchelder, W. H. (1991). Statistical Inference for Multinomial Processing Tree Models. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical Psychology: Current Developments* (pp. 313–335). Springer. https://doi.org/10.1007/978-1-4613-9728-1_18

Rivers, A. M. (2017). The Weapons Identification Task: Recommendations for adequately powered research. *PLOS ONE*, *12*(6), e0177857. https://doi.org/10.1371/journal.pone.0177857

Schmidt, O., Erdfelder, E., & Heck, D. W. (in press). Tutorial on Multinomial Processing Tree Modeling: How to Develop, Test, and Extend MPT Models. *Psychological Methods.* Preprint available at https://doi.org/10.31234/osf.io/gh8md

Schmidt, O., Erdfelder, E., Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000561

Schneider, W., & Shiffrin, R. M. (1977). Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention. *Psychological Review*, *1*(84), 1–66.

Sherman, J. W. (2006). On Building a Better Process Model: It's Not Only How Many, but Which Ones and by Which Means? *Psychological Inquiry*, *17*(3), 173–184.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 485–493. https://doi.org/10.1111/rssb.12062

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*(10), 1332-1345.

Swencionis, J. K., & Goff, P. A. (2017). The psychological science of racial bias and policing. *Psychology, Public Policy, and Law*, *23*(4), 398–409. https://doi.org/10.1037/law0000130

Todd, A. R., Johnson, D. J., Lassetter, B., Neel, R., Simpson, A. J., & Cesario, J. (2021). Category salience and racial bias in weapon identification: A diffusion modeling approach. *Journal of personality and social psychology*, *120*(3), 672–693. https://doi.org/10.1037/pspi0000279

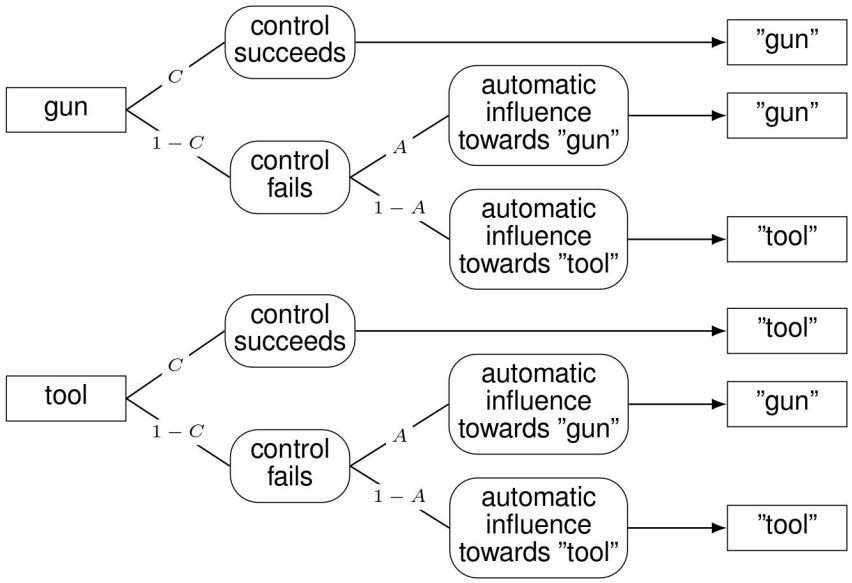
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Volpert-Esmond, H. I., Scherer, L. D., & Bartholow, B. D. (2020). Dissociating automatic associations: Comparing two implicit measurements of race bias. *European journal of social psychology*, *50*(4), 876-888. https://doi.org/10.1002/ejsp.2655

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Figure 1

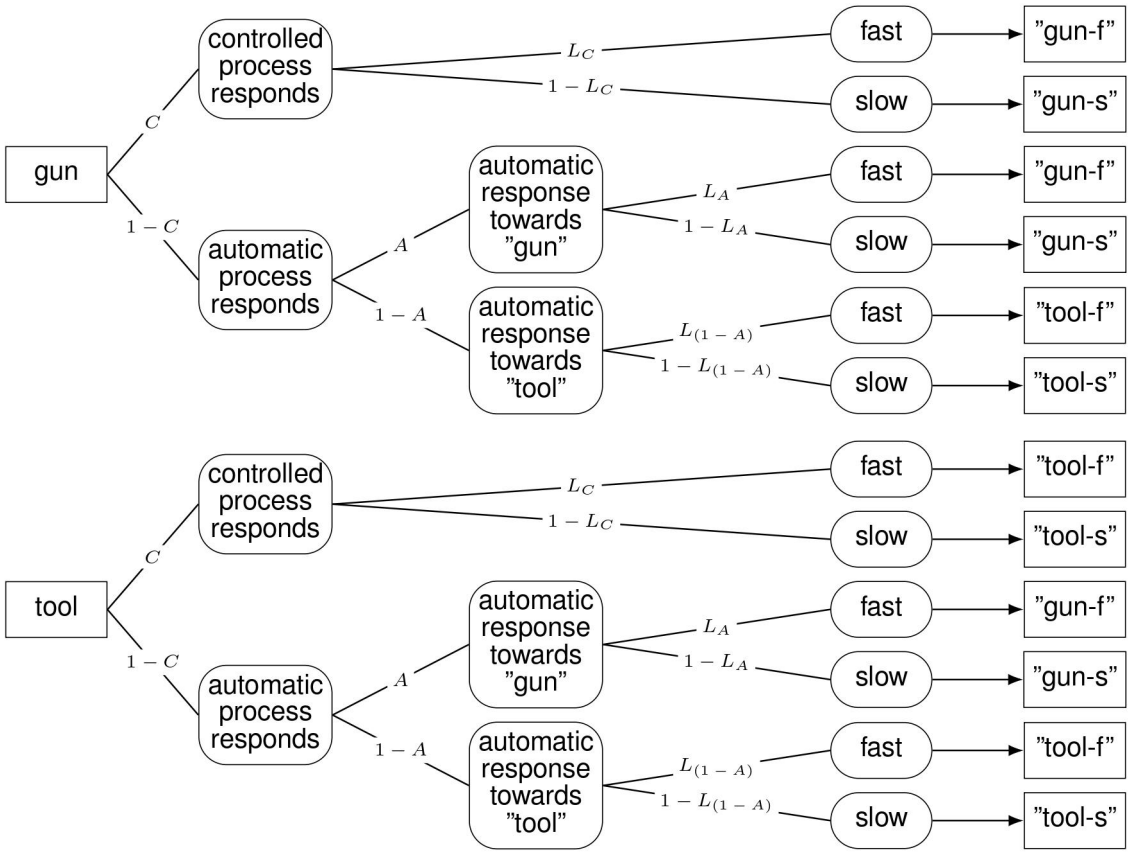
Generic PDP model for the Weapon Identification Task.



Note. Parameters C and A denote probabilities of response determination by a controlled process and an automatic process, respectively. Note that A is conditional on a failure of the controlled process, that is, A represents the conditional probability of response determination by an automatic process given controlled process failure. Each of the parameters C and A may depend on whether the preceding prime is a Black or a White male face.

Figure 2

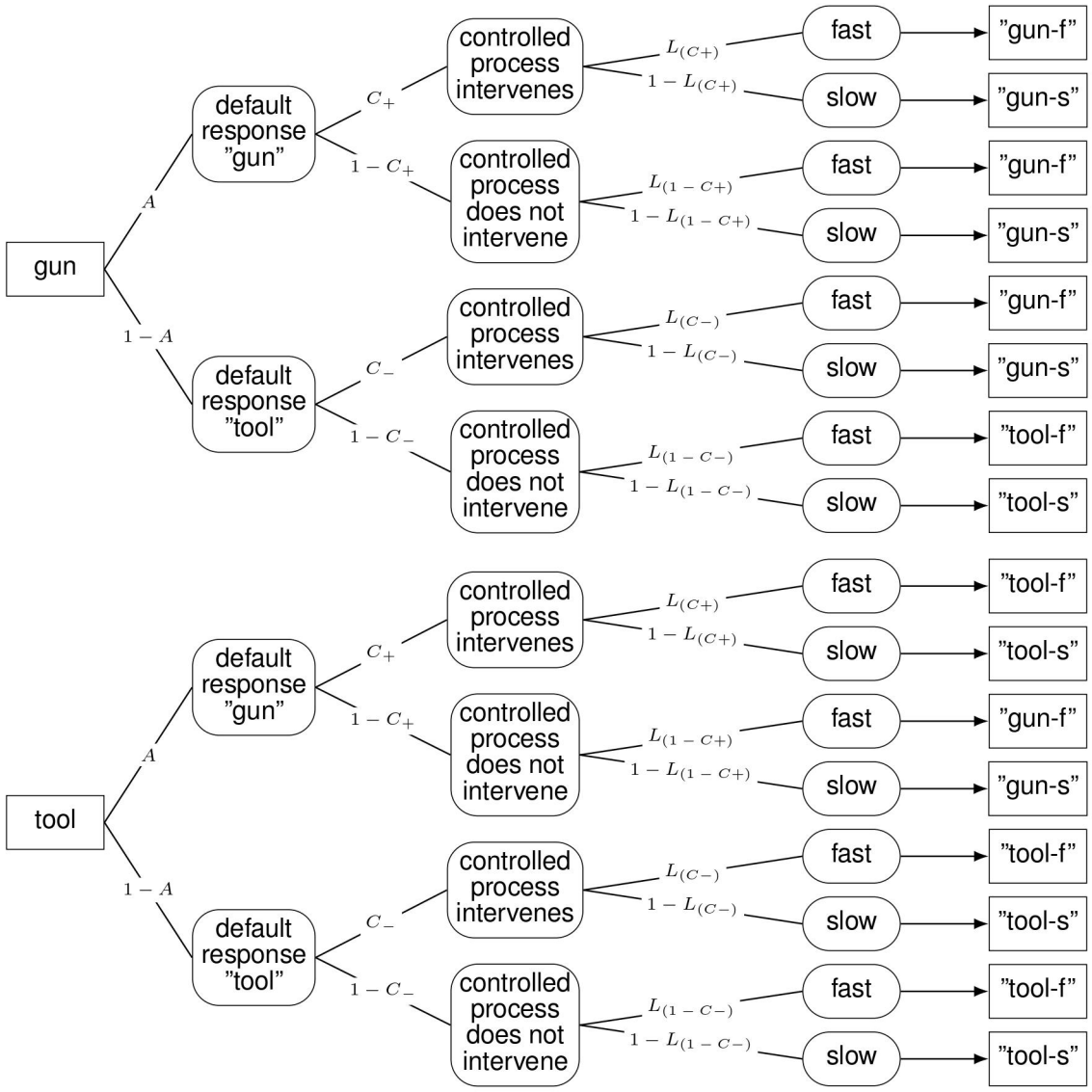
MPT-RT model version of the preemptive conflict-resolution model M_{PCRM}.



Note. The controlled process parameter C and the automatic process parameter A may vary between prime races. The latency parameters L_j may vary as a function of prime races, target objects, and the previous process parameter in the branch the latency parameter belongs to (i.e., C, A, or $(1 - A)$). The category labels "-f" and "-s" indicate responses categorized as fast and slow, respectively.

Figure 3

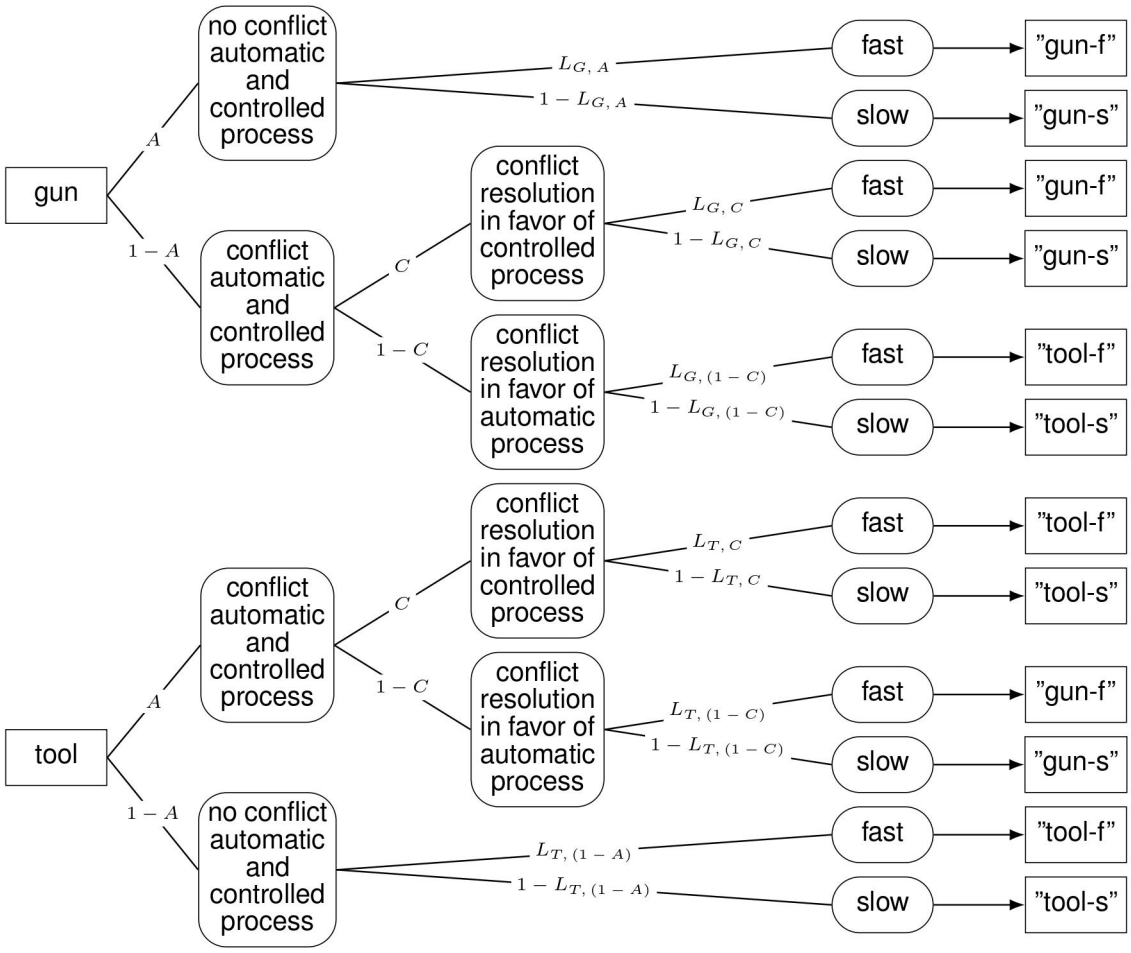
MPT-RT model version of the default interventionist model M_{DIM}.



Note. The controlled process parameter C and the automatic process parameter A may vary between prime races. Subscripts + and − indicate that the parameter is conditional on the default response "*gun*" A vs. "*tool*" $(1 - A)$ suggested by the automatic process, respectively. The controlled process parameter C and the automatic process parameter A may vary between prime races. The latency parameters L_j may vary as a function of prime races, target objects, and the previous process parameter in the branch the latency parameter belongs to (i.e., C or $(1 - C)$). The category labels "-f" and "-s" indicate responses categorized as fast and slow, respectively.

Figure 4

MPT-RT model version of the parallel competitive model M_{PCM}.



Note. The controlled process parameter C and the automatic process parameter A may vary between prime races. The latency parameters L_j may vary as a function of prime races, target objects (G = gun; T = tool) and the previous process parameter in the branch the latency parameter belongs to (i.e., C, $(1 - C)$, A, or $(1 - A)$). The category labels "-f" and "-s" indicate responses categorized as fast and slow, respectively.

Table 1

Overview of key assumptions for cognitive process models of the Weapon Identification Task.

Model	Process Structure	Expected Latency Parameter Pattern
M_{PCRM}: Preemptive conflict-resolution model	An initial preemptive decision results in a controlled process branch with slow responses or in automatic process branches with fast responses.	$L_A = L_{(1-A)} > L_C$
M_{DIM}: Default interventionist model	The default automatic process results in a fast response unless the controlled process corrects the response, resulting in a slow correct response.	$L_{(1-C+)} = L_{(1-C-)} > L_{C+} = L_{C-}$
M_{PCM}: Parallel competitive model	Controlled and automatic processes operate in parallel and result in faster congruent response branches compared to slower conflicting response branches.	$L_{G,A} > L_{G,C} = L_{G,(1-C)}$; $L_{T,(1-A)} > L_{T,C} = L_{T,(1-C)}$
M_{GM}: Guessing model	Successful discrimination results in fast responses, discrimination failure results in additional response time due to guessing.	$L_A = L_{(1-A)} < L_C$

Note. Latency parameters L_j may vary between prime races, target objects (i.e., G = gun; T = tool), and the previous process parameter in the branch the latency parameter belongs to (i.e., C, $(1 - C)$, A, or $(1 - A)$). Subscripts + and − indicate that the parameter is conditional on A and $(1 - A)$, respectively. Concerning the latency patterns, larger L_j parameters indicate faster relative latencies.

Table 2

Characteristics of the eight WIT data sets used in the present analyses.

Article	Study	N_{sample}	% of Caucasian participants	N_{trials}	Prime duration [ms]	Response time limit [ms]	Subset of the data for reanalysis
Amon & Holden (2016)	1a	32	83 % (93 % non-Black)	1,100	200	1,000	only control group
Amon & Holden (2016)	1c	32	83 % (93 % non-Black)	1,100	300	500	only control group
Correll (2008)	2	24	not reported	200	200	1,000	only control group
Lambert et al. (2003)	2	22	all non-black	384	200	550	only private response group
Madurski & LeBel (2015)	1a	48	62 % (98 % non-Black)	200	200	1,000	only control group
Madurski & LeBel (2015)	1b	50	60 % (98 % non-Black)	200	200	1,000	only control groups
Rivers (2017)	1a	40	17 % (98 % non-Black)	216	200	500	neutral faces excluded
Rivers (2017)	1b	42	17 % (98 % non-Black)	216	200	1,000	neutral faces excluded

Note. For the data from Study 2 of Lambert et al. (2003) individual accuracy and response-time data were available for only 22 of 61 participants.

Table 3

Fit statistics for the four process models

Article	Study	M_{PCRM}		M_{DIM}		M_{PCM}		M_{GM}	
		p(T1)	*p(T2)*	*p(T1)*	*p(T2)*	*p(T1)*	*p(T2)*	*p(T1)*	*p(T2)*
AH	1a	.232	.006	.218	.005	<.001	<.001	<.001	<.001
AH	1c	.503	.370	.500	.374	<.001	<.001	<.001	<.001
C	2	.451	.425	.447	.415	<.001	.008	<.001	.008
L	2	.458	.200	.462	.200	<.001	<.001	<.001	<.001
ML	1a	.510	.245	.512	.239	<.001	<.001	<.001	<.001
ML	1b	.525	.194	.521	.177	<.001	<.001	<.001	<.001
R	1a	.528	.338	.519	.348	<.001	<.001	<.001	<.001
R	1b	.507	.236	.502	.246	<.001	.003	<.001	.007

Note. References are abbreviated as follows: AH = Amon & Holden (2016), C = Correll (2008), L = Lambert et al. (2003), ML = Madurski & LeBel (2015), R = Rivers (2017). Models are abbreviated as follows: M_{PCRM} = preemptive conflict-resolution model, M_{DIM} = default interventionist model, M_{PCM} = parallel competitive model, M_{GM} = guessing model.

Table 4

Fit statistics for process model comparison

Article	Study	M_{Basic}		M_{Pars}		Comparison	
		p(T1)	*p(T2)*	*p(T1)*	*p(T2)*	ΔDIC	ΔWAIC
AH	1a	.498 [.001]	.451 [.002]	.037 [.001]	.557 [.002]	0.36 [0.10]	61.91 [0.23]
AH	1c	.492 [.001]	.354 [.002]	.541 [.002]	.379 [.001]	-37.18 [0.08]	-9.77 [0.17]
C	2	.487 [.002]	.357 [.001]	.594 [.001]	.230 [.001]	-16.88 [0.09]	28.45 [0.12]
L	2	.485 [.002]	.232 [.001]	.144 [.002]	.336 [.002]	-23.09 [0.09]	-13.08 [0.10]
ML	1a	.498 [.002]	.220 [.002]	.370 [.001]	.381 [.002]	-32.40 [0.20]	0.29 [0.20]
ML	1b	.480 [.002]	.163 [.002]	.174 [.001]	.086 [.001]	-5.35 [0.14]	20.63 [0.18]
R	1a	.519 [.001]	.370 [.002]	.347 [.001]	.084 [.001]	-0.69 [0.17]	17.05 [0.11]
R	1b	.473 [.002]	.244 [.002]	<.001 [.001]	.437 [.002]	1.52 [0.21]	30.49 [0.17]

Note. *p(T1)*, *p(T2)*, ΔDIC, and ΔWAIC are means with standard errors in brackets across ten independent estimations. Negative values for ΔDIC and ΔWAIC indicate a preference for model M_{Pars} over model M_{Basic}, References are abbreviated as follows: AH = Amon & Holden (2016), C = Correll (2008), L = Lambert et al. (2003), ML = Madurski & LeBel (2015), R = Rivers (2017). Models are abbreviated as follows: M_{Basic} = basic model, M_{Pars} = parsimonious model.

Table 5

Fit statistics and model selection measures for the extended model versions of PCRM and DIM

Article	Study	$M_{PCRM,ext}$		$M_{DIM,ext}$		Comparison	
		p(T1)	*p(T2)*	*p(T1)*	*p(T2)*	ΔDIC	ΔWAIC
AH	1a	.004 [.001]	.031 [.001]	.004 [.001]	.031 [.001]	-10.12 [0.20]	-21.62 [0.23]
AH	1c	.548 [.001]	.408 [.001]	.520 [.002]	.388 [.001]	2.71 [0.11]	-0.14 [0.17]
C	2	.618 [.001]	.233 [.001]	.519 [.002]	.379 [.001]	-3.38 [0.08]	-3.20 [0.15]
L	2	.146 [.001]	.330 [.001]	.209 [.001]	.460 [.002]	-10.31 [0.10]	-13.08 [0.10]
ML	1a	.329 [.001]	.336 [.002]	.337 [.001]	.371 [.002]	-1.35 [0.11]	-9.31 [0.20]
ML	1b	.186 [.001]	.087 [.001]	.349 [.002]	.157 [.001]	-8.38 [0.16]	-17.94 [0.18]
R	1a	.346 [.001]	.082 [.001]	.496 [.002]	.309 [.001]	-8.26 [0.11]	-16.84 [0.14]
R	1b	<.001 [.001]	.432 [.001]	.230 [.002]	.440 [.002]	-18.59 [0.16]	-27.91 [0.20]

Note. *p(T1)*, *p(T2)*, ΔDIC and ΔWAIC are means with standard errors in brackets across ten independent estimations. Negative values for ΔDIC and ΔWAIC indicate a preference for model $M_{DIM,ext}$ relative to $M_{PCRM,ext}$. References are abbreviated as follows: AH = Amon & Holden (2016), C = Correll(2008), L = Lambert et al. (2003), ML = Madurski & LeBel (2015), R = Rivers (2017). Models are abbreviated as follows: $M_{PCRM,ext}$ = extended version of preemptive conflict resolution model, $M_{DIM,ext}$ = extended version of default interventionist model.

Towards a Process-level Understanding of Correspondence among Implicit Measures of Racial Bias

Ruben Laukenmann[1,2] & Jimmy Calanchini[2]

[1] University of Mannheim

[2] University of California, Riverside

Author Note

Ruben Laukenmann https://orcid.org/0000-0002-4780-4845

Jimmy Calanchini https://orcid.org/0000-0003-1959-143X

Manuscript preparation was supported by the Research Training Group "Statistical Modeling in Psychology", funded by the Deutsche Forschungsgemeinschaft (GRK 2277), by a Doctoral Research Fellowship from the Stiftung der Deutschen Wirtschaft (sdw) gGmbH, funded by the German Federal Ministry of Education and Research, and by a one-year research grant for doctoral candidates funded by the German Academic Exchange Service (DAAD) awarded to Ruben Laukenmann.

The authors are very grateful to C. E. Phills and A. M. Rivers for sharing the stimuli material we used for our studies.

Correspondence concerning this article should be addressed to Ruben Laukenmann, Cognition and Individual Differences Lab, University of Mannheim, D-68159 Mannheim, Germany (email: ruben.laukenmann@psychologie.uni-mannheim.de) or Jimmy Calanchini, Department of Psychology, University of California Riverside, Riverside, CA 92521 (email: jimmy.calanchini@ucr.edu)

Online supplementary material is available at https://osf.io/2whze/?view_only=f87cf1d5308a48d5a5553d64b216133c

This research was approved by the Institutional Review Board of the University of California, Riverside.

We have no conflict of interest to declare.

Abstract

Several implicit bias measures aim to assess racial threat stereotypes. However, previous research often reveals relatively poor correspondence among three implicit measures – the Weapon Identification Task (WIT), the First-Person Shooter Task (FPST), and the Implicit Association Test (IAT) – which is surprising for tasks configured to assess a common construct. Importantly, these measures differ on a wide variety of procedural dimensions (e.g., task instructions, stimulus presentation, response time limits), which may explain low correlations among tasks. In the present research, $N = 372$ participants each completed versions of the WIT, FPST, and IAT that were equated on several procedural dimensions. Process modeling revealed high correlations among tasks for model parameters reflecting controlled and automatic processes. Our findings suggest that all three measures can correspond with one another when procedurally aligned and analyzed with sufficient theoretical precision.

Abstract: 135 words

Keywords: weapon identification task, first person shooter task, implicit association test, multinomial processing tree modeling, process model, racial bias

Word count: 8946 (excluding references, tables, figures, and footnotes)

Towards a Process-level Understanding of Correspondence among Implicit Measures of Racial Bias

Introduction

Implicit measures[1] were initially developed to assess the strength of mental associations stored in memory by minimizing the influence of other mental processes that would constrain their expression. Tasks such as the Weapon Identification Task (WIT; Payne, 2001), First-Person Shooter Task (FPST; Correll et al., 2002), and the Implicit Association Test (IAT; Greenwald et al., 1998) are often configured to assess implicit racial bias operationalized as attitudes (i.e., evaluations) or stereotypes (i.e., traits) associated with different racial groups[2]. However, previous research has revealed relatively weak to no correlations between different measures of implicit racial bias (Cunningham et al., 2001; Glaser & Knowles, 2008; Ito et al., 2015; Payne, 2005; Volpert-Esmond et al., 2020), which is surprising for measures that are configured to assess a common construct.

One possible explanation for low correlations among implicit racial bias measures is that they actually assess conceptually distinct constructs. For example, we should not expect – nor do we find – that an implicit measure configured to assess racial attitudes to correspond strongly to an implicit measure configured to assess racial stereotypes (e.g., Amodio & Devine, 2006; Glaser & Knowles, 2008; Payne, 2005; Volpert-Esmond et al., 2020). However, relatively small correlations also emerge among different implicit bias measures configured to assess the same construct (e.g., Bar-Anan & Nosek, 2014; Cunningham et al., 2001; Ito et al., 2015; Olson & Fazio, 2003). Small correlations among conceptually analogous measures necessarily calls into question the validity of implicit measures, and of the construct of implicit bias more generally (Schimmack, 2021). Nevertheless, even when implicit bias measures are configured to assess the same construct, a wide variety of procedural differences among measures remain, such as differences in stimuli, response time limits, number of trials, task instructions, and stimuli presentation procedures. Task procedures necessarily determine which processes influence responses on implicit measures (Gawronski et al., 2010), so procedural differences may translate into small correlations across implicit measures that otherwise assess the same

construct. Thus, in the present research, we assess correspondence across three measures of implicit threat stereotypes – the WIT (Payne, 2001), the FPST (Correll et al., 2002), and the IAT (Greenwald et al., 1998) – that are equated across several procedural dimensions. Moreover, responses on implicit measures are traditionally operationalized in terms of summary statistics, which provide little insight into the cognitive process(es) that contribute to responses (Calanchini, 2020). Consequently, in the present research we apply a variety of formal mathematical models to responses on each measure to provide theoretically precise insight into the processes that produce responses, and their correspondence among processes across measures.

Process Modeling and Implicit Measures of Threat Stereotypes

We are not the first to recognize that procedural differences across implicit bias measures affect correspondence across measures (e.g., Olson & Fazio, 2003), nor are we the first to use formal mathematical modeling to assess correspondence across implicit threat stereotype measures. Indeed, Ito and colleagues (2015) investigated the contributions of two qualitatively distinct types of cognitive processes on the WIT, FPST, and IAT. Estimates of controlled processing correlated moderately-to-strongly across measures ($r = .29 - .61$) but estimates of racial bias based on automatic processing correlated weakly across measures ($r = .02 - .16$).

To the extent that implicit measures were developed to primarily assess automatic mental processes (Fazio et al., 1995; Greenwald et al., 1998), the small magnitude of racial bias estimate correlations calls into question whether implicit bias measures configured to assess threat stereotypes reflect a common construct. However, and importantly, Ito and colleagues (2015) relied on the standard versions of each measure as they were originally published. This approach certainly has ecological validity, in terms of correspondence with how the measures are traditionally used by researchers. But as we review below, the three measures differ from one another on many dimensions – which in turn confounds any strong interpretations of the small correlations in automatic processing between tasks.

Overview of Standard Versions of Implicit Measures

Weapon Identification Task

The WIT is a sequential priming paradigm introduced by Payne (2001). Over a series of trials, participants view a prime image (i.e., Black or White male face) quickly followed by a target image (i.e., gun or tool). On each trial, participants' task is to identify quickly and accurately the target while disregarding the prime (Payne, 2001; Rivers, 2017).

First-Person Shooter Task

The FPST is a simplified videogame simulation introduced by Correll et al. (2002). Over a series of trials, participants view a naturalistic scene (e.g., a park) in which a person (i.e., Black or White) appears who is either armed (e.g., holding a gun) or unarmed (e.g., holding a cell phone). On each trial, participants' task is to quickly and accurately decide whether to "shoot" or "don't shoot". Typically, target persons are presented as full body photographs with varying postures against different background scenes, with changing target onset times and screen positions (Correll et al., 2002; 2007; 2015). However, simplified versions of the FPST present only pictures of faces with the target object (e.g., gun, cell phone) superimposed on or positioned next to the face (Correll et al., 2014; Plant et al., 2005; Unkelbach et al., 2008).

Implicit Association Test

The IAT is a dual-categorization task introduced by Greenwald et al. (1998). Over a series of trials, participants view two stimulus types – typically target groups (e.g., pictures of Black and White male faces) and attributes (e.g., words referring to threat or safety) – presented one at a time. In some blocks of trials, one target type shares a response key with one attribute type (e.g., Black/threat) and the other target type shares a response key with the other attribute type (e.g., White/safety). However, in other blocks of trials, the key pairings are reversed (e.g., Black/safety, White/threat). On each trial, participants' task is to categorize the presented stimulus quickly and accurately.

Similarities and differences among standard versions of implicit measures

All three of these measures – the WIT, FPST, and the IAT – have been used to assess implicit racial bias operationalized in terms of threat stereotypes. Moreover, all three tasks share a common feature in that they require participants to make a categorization decision (e.g., whether the stimulus is a gun or tool). However, as Table 1 illustrates, the three measures diverge on several procedural characteristics. Perhaps the biggest difference across tasks is the task structure itself. The WIT presents stimuli sequentially and explicitly instructs participants to respond to target objects and disregard face primes. The FPST presents stimuli concurrently, and participants are instructed to "shoot" an armed person and "don't shoot" an unarmed person. The IAT presents stimuli serially, such that participants must attend and categorize both types of stimuli. Additionally, the standard version of each of these measures differs in the number of critical trials, the presence and length of the response time limit, task instructions, and stimulus material.

Insert Table 1 about here

Methodological differences necessarily introduce variance between measures, which in turn may obscure relationships among measures of implicit racial bias (Fazio et al., 1995; Greenwald & Lai, 2020; Mekawi & Bresin, 2015; Payne, 2005; Olson & Fazio, 2003; Volpert-Esmond et al., 2020). Differences in the stimuli presentation traditionally used in each task provide a straightforward illustration of this point. For example, the WIT presents the target object clearly visible in the center of the screen without any distracting cues, and at approximately the same size as the face primes are presented. In contrast, the FPST presents the target object held in the hands of a person who is pictured in a variety of scenes, and thus represent only a small feature in a larger context. The IAT differs from both of these paradigms, and typically does not use images of target objects but, instead, represents attributes like safety versus danger using words (Ito et al., 2015). Previous research has demonstrated that stimuli that are intended to reflect different operationalizations of the same construct (e.g., words versus pictures) can nevertheless activate different sets (or subsets) of

associations (Foroni & Bel-Bahar, 2010; Rosch, 1975). Thus, low correspondence among measures may reflect different kinds of stimuli presentations activating different subsets of threat associations. Moreover, structural differences in stimuli presentation (i.e., sequential, concurrent, serial) across measures may also contribute to their low correspondence. For example, previous research investigating relationships between implicit measures and executive functions has shown that the WIT and FPST are both related to inhibition (Ito et al., 2015; Payne, 2005), whereas the IAT is related to updating and task switching (Ito et al., 2015; Klauer et al., 2010). To the extent that all three implicit measures are configured to assess a common construct (i.e., threat stereotypes), the influence of different executive functions would seem to be related to procedural differences across measures.

Though all three measures differ on a variety of procedural dimensions, we propose that task structure (i.e., sequential, concurrent, serial) is the most crucial difference. Table 1 describes the procedures of each measure as it was initially proposed, but nevertheless each of these measures has been implemented with procedures that differ from the ones described here. For example, the WIT has been used with additional neutral face outlines (Rivers, 2017), and faces depicting various demographic information like race, gender, and age (Stein et al., 2023; Thiem et al., 2019; Todd et al., 2021). The FPST exists in several different adaptations presenting full body images of suspects holding the target object in their hand (Correll et al., 2002; Payne & Correll, 2020), target objects being superimposed on the forehead of a face (Plant et al., 2005), or target objects presented on the left or right side of a person's face (Unkelbach et al., 2008; Unkelbach et al., 2009). Furthermore, IATs have been developed that implement response deadlines, different numbers of trials, and different block structures (Calanchini et al., 2021; Meissner & Rothermund, 2013; Sriram & Greenwald, 2009). However, across all these implementations, task structure remains unchanged from the original version. The WIT present stimuli sequentially, the FPST presents stimuli concurrently (Payne & Correll, 2020)[3], and the IAT presents stimuli serially.

Implicit Measures Reflect the Contributions of Multiple Cognitive Processes

Despite procedural differences among implicit measures, they were all designed with the same goal of assessing mental associations between target groups and attributes (e.g., race and threat) by

minimizing the contributions of other processes that would constrain the expression of associations. To be sure, compared to analogous direct, self-report, or explicit measures, implicit measures minimize the contributions of motivations and biases that would modify responses in a socially desirable direction. However, responses on implicit measures do not solely reflect the contributions of mental associations. Instead, a growing body of research uses multinomial processing tree models (MPTs; Riefer & Batchelder, 1988) to disentangle the joint contributions of multiple cognitive processes to responses on implicit measures. MPTs are tailored to specific experimental paradigms that provide frequency data (e.g., number of correct and incorrect responses), and specify the number, nature, and composition of cognitive processes thought to be involved in the paradigm (for reviews see Calanchini, 2020; Erdfelder et al., 2009; Hütter & Klauer, 2016). In creating MPT models, researchers must make theoretically grounded decisions about the specific way multiple cognitive processes produce responses in each task condition. The proposed cognitive processes are represented by model parameters, and their proposed interplay can be illustrated in a processing tree that consists of a root with multiple branches, with each branch corresponding to the success or failure of a process or series of processes. Each process is conditional upon any processes that precede it in a given tree branch. The model estimates parameter values that most closely approximate participants' observed responses across task conditions, and these parameter estimates are interpreted as probabilities that each cognitive process influenced participants' responses on the paradigm.

MPT models are mathematical instantiations of psychological theory packaged in a well-defined form. A variety of theories have been proposed to account for the cognitive processes that underlie responses on implicit measures and, accordingly, a variety of MPTs have been proposed to disentangle the influence of multiple processes to implicit measures of racial bias. In the following we provide an overview of the seven MPT models that we will apply in the present research. We describe each model in the context of the WIT, but the assumptions of the models also hold true for the FPST and IAT.

Process Dissociation Procedure

The Process Dissociation Procedure (PDP; Jacoby, 1991; Payne, 2001) is depicted in Figure 1 (upper panel). This model consists of a Controlled (C) and an Automatic (A) process parameter. The Controlled process parameter represents any process(es) that result in a correct response, including but not limited to general accuracy in responding based on successful target discrimination, as well as conflict monitoring/resolution when the automatic processes would produce a response that conflicts with the correct response (Klauer & Voss, 2008; Laukenmann et al., 2023). The Automatic process parameter represents response tendencies towards producing a "gun" versus "tool" response, which include automatic threat stereotype associations. Success of the Controlled process (C) will always produce the correct response (+). The Automatic process can only produce a response in the absence of influence from the Controlled process ($1 - C$), with the probability (A) representing a "gun" response and with the counter-probability ($1 - A$) representing a "tool" response. When the target stimulus is a gun, (A) produces a correct response (+), and ($1 - A$) produces an incorrect response (–). In contrast, when the target stimulus is a tool, (A) produces an incorrect response (–), and ($1 - A$) produces a correct response (+). In the PDP, controlled processing dominates automaticity, such that the influence of automatic processing is irrelevant whenever the controlled process succeeds.

PDP with guessing

The Process Dissociation Procedure with guessing (PDP+G, Bishara & Payne, 2009) is an extension of the PDP, and is depicted in Figure 1 (lower panel). Like the PDP, the PDP+G consists of a Controlled (C) and an Automatic (A) parameter but adds a Guessing (G) process parameter. The Controlled process parameter represents accuracy in responding and conflict monitoring/resolution that will always produce a correct response, and the Automatic process parameter represents the activation of automatic threat stereotype associations that can only influence responses in the absence of influence from the Controlled process. The Guessing process parameter in the PDP+G represents a general, stereotype independent, response tendency that can only influence responses in the absence of influence from either the Controlled or Automatic process. The term "guessing" is a bit of a misnomer, as it does not necessarily reflect random responding; instead, this parameter is a catch-all

that accounts for any processes that influence responses and are not already reflected in the Controlled and Automatic process parameters. In the absence of influence from the Controlled process ($1 - C$), the Automatic process either influences the response towards "gun" (A) or does not influence the response ($1 - A$). If neither the Controlled ($1 - C$) nor the Automatic ($1 - A$) process influences the response, the response is guessed towards "gun" (G) or "tool" ($1 - G$). Whereas the meaning of the Controlled process is the same in the PDP and in the PDP+G, the meaning of the Automatic process is changed: because the Guessing parameter represents any other processes that would produce a "gun" versus "tool" response, the Automatic process parameter primarily represents the influence of automatic stereotype association towards "gun".

Insert Figure 1 about here

Stroop Model

The Stroop model is depicted in Figure 2 (upper panel). Lindsay and Jacoby (1994) proposed this model for the Stroop task (Stroop, 1935) as an alternative to the PDP model. Like the PDP, the Stroop model consists of a Controlled (C) process parameter that represents accuracy in responding and conflict monitoring/resolution, and an Automatic (A) process parameter that represents processes that produce "gun" versus "tool" responses, including the activation of automatic threat stereotype associations. However, in contrast to the PDP model, in the Stroop model the Automatic process is assumed to dominate responses. If the Automatic process succeeds (A), it influences the response towards "gun" leading to a correct response (+) in gun-target trials and an incorrect response in tool-trials (–)[4]. However, if the Automatic process has no influence ($1 - A$), the success of the Controlled process (C) produces a correct response (+), and the failure of the Controlled process ($1 - C$) produces an incorrect response (–).

Stroop Model with Guessing

The Stroop model with Guessing (Stroop+G, Bishara & Payne, 2009) is an extension of the Stroop model, and is depicted in Figure 2 (lower panel). Like the PDP+G model, the Stroop+G

model consists of a Controlled (C), an Automatic (A), and a Guessing (G) process parameter. Like in the PDP+G model, in the Stroop+G model the Controlled process parameter represents accuracy in responding and conflict monitoring/resolution for interfering processes, the Automatic process parameter represents the activation of automatic threat stereotype associations, and the Guessing process parameter represents a general, stereotype independent, response tendency that can only influence responses in the absence of influence from either the Controlled or Automatic process. If the Automatic process succeeds (A), it influences the response towards "gun", leading to a correct response (+) in gun-target trials, and an incorrect response in tool-trials (–). In the absence of influence from the Automatic process ($1 - A$), the Controlled process either (C) produces a correct response (+) or does not influence the response ($1 - C$). If neither the Automatic ($1 - A$) nor the Controlled ($1 - C$) processes influences the response, the response is guessed towards "gun" (G) or "tool" ($1 - G$). Like the Stroop model, in the Stroop+G model automaticity dominates controlled processing and guessing, such that the influence of controlled processing and guessing is irrelevant whenever the automatic process succeeds.

Insert Figure 2 about here

Quad Model

The Quad model (Conrey et al., 2005) is depicted in Figure 3. The Quad model consists of four parameters: Detection (D), Automatic association activation (AC), Overcoming bias (OB), and Guessing (G). The Detection process parameter represents accuracy in responding based on successful target discrimination. The Automatic process parameter represents the activation of automatic threat stereotype associations for Black males and non-threat stereotype associations for White males. The Detection and Automatic parameters of the Quad model are conceptually analogous to the Controlled and Automatic parameters of the PD(+G) and Stroop(+G) models. However, the Quad model differs from the other models in its assumptions about the dominance of either process. Whereas the PD(+G) models assume that the Controlled process will always drive a response if activated, and the

Stroop(+G) models assume that the Automatic process will always drive a response if activated, the Quad model allows for both types of process to potentially drive a response if activated. When Detection and Automatic processes are both activated and would produce conflicting responses (i.e., a correct response versus a stereotype-congruent incorrect response, respectively), the Overcoming Bias process parameter represents successful conflict resolution over the Automatic processes. Like in the PDP+G and Stroop+G models, the Guessing process parameter in the Quad model represents any processes that would produce a general, stereotype independent, response tendency that are not otherwise accounted for by the other model parameters.

If the Automatic process succeeds (AC) and Detection is achieved (D), the automatic influence can be overcome (OB) or not ($1 - OB$). If the automatic influence is overcome (OB), Detection leads to the correct response (+), but if the Automatic process is not overcome ($1 - OB$), the Automatic process drives a response that is correct (+) on "gun" trials but incorrect (-) on "tool" trials. If the Automatic process ($1 - AC$) and detection have no influence ($1 - D$), the response is guessed towards "gun" (G) or "tool" ($1 - G$).

Insert Figure 3 about here

As originally specified by Conrey et al. (2005), the Quad model assumes distinct associations for Black versus White targets: "Black" and "gun", but "White" and "tool". In the following, we refer to this model specification as the *traditional* Quad model. Consequently, in the traditional Quad model, "White" cannot be associated with "gun", or "Black" with "tool". This assumption contrasts with the PDP model, which allows for both directions of association: automatic parameter values greater than .5 reflect "gun" associations and less than .5 reflect "tool" associations. In the present research, we posit that "gun" associations are more theoretically relevant than "tool" associations to racial threat stereotypes. We implement this assumption in the form of a respecified *egalitarian* Quad model, such that both automatic parameters reflect "gun" associations (i.e., Black-"gun" and White-"gun"). Thus, the egalitarian Quad model assumes the same direction for the pattern of associations

but allows for the strength of the activated association to vary by race. For the sake of completeness, we will include both model specifications – the traditional Quad model and the egalitarian Quad model – in the present research.

Stereotype Misperception Task Model

The Stereotype Misperception Task (SMT) model (Krieglmeyer & Sherman, 2012) is depicted in Figure 4. The SMT model consists of four parameters: Detection (*D*), Stereotype Activation (*SAC*), Stereotype Application (*SAP*), and Guessing (*G*). The Detection process parameter represents general accuracy in responding based on successful target discrimination. The Stereotype Activation process parameter represents the activation of automatic threat stereotype associations. The Stereotype Application process parameter represents whether the activated threat stereotype is applied or not for responding. The Guessing process parameter represents a general, stereotype independent, response tendency. If a stereotype is activated (*SAC*), it produces a response tendency towards "gun". This activated stereotype can be applied (*SAP*) which produces a "gun", or it can be corrected (1 – *SAP*) which produces a "tool" response. If no stereotype is activated (1 – *SAC*), detection (*D*) will produce a correct response (+). If no stereotype is activated (1 – *SAC*) and detection is not achieved (1 – *D*), guessing produces a "gun" (*G*) or a "tool" response (1 – *G*).

Insert Figure 4 about here

The Present Research

The aim of this study is to investigate correspondence across three implicit measures – the WIT, FPST, and IAT – configured to assess racial threat stereotypes. Importantly, we retain each measure's traditional structure (i.e., sequential, concurrent, serial) and instructions but, in contrast to previous research, align them on all other procedural dimensions. Specifically, in the present research we used the same stimuli, response time window, and number of trials in all three measures. In doing so, we can investigate the extent to which the small correlations among measures identified in

previous research (Glaser & Knowles, 2008; Ito et al., 2015; Payne, 2005) reflect procedural artifacts versus meaningful differences among the measures.

In the present research, each participant completed all three implicit measures, and we investigated correspondence among measures in two main ways. First, we examined correlations among summary statistics of performance (i.e., difference in accuracy between critical trials). Next, to provide more theoretically precise insight than can be revealed by summary statistics, we applied MPT models to participants' responses on each measure. We examined which MPT models provide good fit to the data from each measure and how the model parameters correlate across measures. Replicating previous research, we expect parameters that reflect controlled processes to correlate moderately-to-strongly between measures (Ito et al., 2015). Additionally, and in contrast to previous research, we expect for parameters that reflect automatic processes and racial bias to correlate moderately-to-strongly between measures because we have aligned stimuli across measures.

We applied all seven MPTs (PDP, PDP+G, Stroop, Stroop+G, traditional Quad, egalitarian Quad, and SMT) to participants' responses on each measure. Given that an MPT model reflects theoretical assumptions specified in equation form, we can quantitatively examine the extent to which the assumptions articulated in each MPT model fit each measure. In comparing models, we will select the best-fitting model (as identified by model-selection indices) for each measure, then investigate correlations across measures between conceptually analogous parameters.

Though our model comparisons will be in part exploratory, we nevertheless can make some predictions about which model might be favored for each measure. We expect the PDP to provide best fit to the WIT because the PDP was initially developed for the WIT (Payne, 2001) and has been widely used with the WIT (Bishara & Payne, 2009; Ito et al., 2015; Klauer & Voss, 2008; Laukenmann et al., 2023; Lambert et al., 2003; Payne et al., 2002). That said, the PDP has not provided unambiguously best fit to the WIT in previous research: for example, Burke (2015) concluded that the Quad model provided superior fit to the WIT than did the PDP. We expect the Quad model to provide best fit (either in its traditional or egalitarian version) to the IAT because it was initially developed for the IAT (Conrey et al., 2005) and has been widely used with the IAT.

However, also the PDP model has been applied before to analyze IAT data (Ito et al., 2015). We have no predictions about which MPT model would provide best fit to the FPST. The FPST is typically analyzed using signal detection modeling (Correll et al., 2002; Mekawi & Bresin, 2015), though the PDP model has been applied to the FPST in previous research (Huntsinger et al., 2009; Ito et al., 2015).

Methods

Participants

In total, 547 undergraduate students at a large, public Southern California university participated for partial course credit. As a prerequisite of the Institutional Review Board at the university where the study took place, participants were excluded from analysis if they chose to reject data inclusion at the end of the study (44 rejected inclusion). Participants were also excluded whose error rate was >50% for any measure (9 WIT; 2 FPST; 5 IAT; 1 in at least two of the measures), or missing trials at a rate 1.5 times the interquartile range above the median (Tukey's criterion) for at least one measure (43 WIT; 26 FPST; 6 IAT; 46 in at least two of the measures).

The final sample comprised 372 participants (age: M_{age} = 19.6, SD_{age} = 2.3; gender: 225 female, 141 male, 5 other, 1 unreported; race: 33 White, 11 Black, 130 Asian, 146 Latino, 48 other, 4 unreported). A post-hoc power analysis for a repeated measurement ANOVA with a sample size of N = 372, and a Type-1 error level of α = .05 (Faul et al., 2009) afforded a test power of at least 1-β > .49 to detect small effects (f = .1 or η_p^2 = .01 in the underlying population, cf. Cohen, 1988). If effects are at least of medium size (f = .25 or η_p^2 = .06 in the underlying population, cf. Cohen, 1988), the power increases to 1-β > .99 under otherwise identical conditions. A post-hoc power analysis for a bivariate correlation and a two-tailed Type-1 error level of α = .05 afforded a test power of at least 1-β = .49 to detect a small correlation (r = .10) and increased to a test power of at least 1- β > .97 to detect a correlation of r =. 20 given identical conditions. Overall, our study is sufficiently powered.

Procedure

Data collection was conducted online. After providing consent, participants completed the three implicit measures in random order. Then, participants completed a basic demographic

questionnaire (age, gender, ethnicity) and were asked what they thought the study was about. Participants were thanked and debriefed at the end. Across all measures, the response keys D and L were counterbalanced between participants, and the response keys corresponding to gun and tool targets were held constant within participants.

Weapon Identification Task. Participants completed an adapted version of the WIT (Payne, 2001; Rivers, 2017). Participants were instructed to identify as quickly and accurately as possible a target object (i.e., gun, tool) preceded by a face image as prime. In addition to the Black and White male faces that are traditionally used as primes in the WIT, the adapted WIT also included an outline of a face as neutral prime to provide sufficient degrees of freedom for MPT modeling. In each trial, participants were presented with a fixation cross (500 ms), a face prime (200 ms), a target object (200 ms), a pattern mask (500 ms)[5], and a feedback screen (1000 ms), each presented in the center of the screen. On practice trials, participants received the following feedback: 'correct!', 'false!', 'too slow!'. Slow responses were operationalized as responses made 700 ms or more after target object onset. On experimental trials, participants only received feedback if their response was too slow. Participants' response latency was recorded, even if it exceeded the 700 ms limit. Participants first completed 20 practice trials containing only the neutral face outlines as primes, half of which were paired with a gun target and half of which were paired with a tool target. Next, participants completed 240 experimental trials with 80 trials for each prime race by target object combination in random order. Participants had two self-paced breaks after 80 and 160 trials.

First-Person Shooter Task. Participants completed an adapted version of the FPST (Correll et al., 2002, 2014; Unkelbach et al., 2008). Participants were instructed to "shoot" or "don't shoot" as quickly and accurately as possible an image of a Black male face, a White male face, or the outline of a face as neutral image, that was paired with either a gun ("armed") or tool ("unarmed") target object. The target object was displayed on either the left or right side of the face with its handle pointing towards the face. In each trial, participants were presented with a fixation cross (500 ms), a face image and target object presented simultaneously at one of nine random positions on the screen (700 ms), and a feedback screen (1000 ms). On practice trials, participants received the

following feedback: 'correct!', 'false!', 'too slow!'. Slow responses were operationalized as response made 700 ms or more after the onset of the target object with the face image. On experimental trials, participants only received feedback if their response was too slow. Participants first completed 20 practice trials containing only neutral face outlines as face images, half of which were paired with a gun target and half of which were paired with a tool target. Next, participants completed 240 experimental trials with 80 trials for each face race by target object combination in random order. Participants had two self-paced breaks after 80 and 160 trials.

Implicit Association Test. Participants completed an adapted version of the IAT (Greenwald et al., 1998). Participants were instructed to categorize as quickly and accurately as possible target objects (a gun or a tool) and face images (a Black male face, a White male face, or a neutral outline of a face). In each trial, participants were presented with a fixation cross (500 ms), a target object or face image (700 ms), and a feedback screen (1000 ms), each presented in the center of the screen. On practice trials, participants received the following feedback: 'correct!', 'false!', 'too slow!'. Slow responses were operationalized as responses made 700 ms or more after the onset of the target object or face image. On experimental trials, participants only received feedback if their response was too slow. In each trial, key assignments for the categories (target objects: "gun"/"tool"; race categories: "Black"/"White"/"neutral") were continuously displayed on the left or right lower corner on screen. In total, the IAT consisted of 13 blocks, seven practice blocks with 20 trials and six experimental blocks with 40 trials each. The first practice block consisted of learning the assignment of the target objects *gun* and *tool* to the left or right response key. Next, each of the three combinations of face image category pairings (Black male vs. White male; Black male vs. neutral outline; White male vs. neutral outline) were presented in a grouping of four blocks. These four blocks each consisted of: a practice block learning the assignment of the race categories to the left and right response keys, an experimental block combining target objects and face images, a practice block with reversed assignment of the race categories to the response keys, and an experimental block combining target objects and face images with reversed key assignment for the face images. The order of key

pairings for target objects with race categories was randomized within each race category pairing. The four blocks of each race category pairing were presented in a random order for each participant.

Materials

Each implicit measure included the same stimulus material taken from Rivers (2017) and Phills et al. (2011) and were displayed with a 300 × 300-pixel resolution. Face primes consisted of 24 Black male faces, 24 White male faces, and one neutral image of the outline of a face. Target objects consisted of 5 drawings of guns and of 5 tools presented horizontally.

Results

Data pre-processing

Prior to all analyses, we excluded trials with latencies <100 ms and >1700 ms, resulting in exclusion of 1.05% of trials. The implicit measures were presented in random order for each participant to rule out order effects, but we do not include task order in our analysis.

Error rate analysis[6]

Figure 5 shows the error rates of all three measures by race and target. A 3 (measure) × 3 (race) × 2 (target) mixed analysis of variance (ANOVA) with Greenhouse-Geyser correction on the error rates yielded a main effect of measure $F(1.68, 622.45) = 60.75$, $p < .001$, $\eta_p^2 = .141$, which indicates that error rates varied across measures. To investigate these differences, we used paired *t*-test comparisons. The WIT had a significantly higher error rate ($M = 14.0\%$, $SD = 9.4$) than the FPST ($M = 11.1\%$, $SD = 6.9$), $t(371) = 6.37$, $p < .001$, $d_Z = 0.25$, and a higher error rate than the IAT ($M = 9.7\%$, $SD = 5.2$), $t(371) = 9.92$, $p < .001$, $d_Z = 0.38$. The FPST had a significantly higher error rate than the IAT, $t(371) = 5.01$, $p < .001$, $d_Z = 0.17$.

The ANOVA also yielded the expected race × target interaction, $F(1.96, 726.08) = 63.08$, $p < .001$, $\eta_p^2 = .145$, indicating that error rates for each target type varied as a function of race. A three-way interaction also emerged, $F(3.71, 1374.92) = 3.53$, $p = .009$, $\eta_p^2 = .009$, indicating that the racial bias effect varied across measures. Nevertheless, analyzing each measure separately, the race by target interaction remained significant for all measures: WIT ($F(1.86, 690.93) = 26.53$, $p < .001$, $\eta_p^2 =$

.067), FPST (F(1.98, 735.68) = 17.56, p <.001, η_p^2 = .045), and IAT (F(1.98, 735.79) = 38.13, p <.001, η_p^2 = .093).

Insert Figure 5 about here

To investigate the correspondence between measures we analyzed the correlations between summary statistics of accuracy as proxies for racial bias. To do so, we calculated the difference between errors for gun and tool targets for each face type. Specifically, we calculated Black versus White accuracy bias as: (errors(tool|Black) – errors (gun|Black)) – (errors (tool|White) – errors (gun|White)). Conceptually replicating previous research, accuracy bias estimates correlated significantly between the WIT and FPST (r = .18, p < .001), but not between the WIT and IAT (r = .06, p = .27) nor between the FPST and IAT (r = .02, p = .77).

Multinomial Process Tree Model estimation

Modeling procedure. We conducted MPT modeling to investigate which model fits and explains the data best for each implicit measure. We used the hierarchical, latent-trait model of Klauer (2010) as a framework for all analyses.

To assess goodness-of-fit, we used Bayesian posterior predictive p-values corresponding to the test statistics T_1 and T_2 (Klauer, 2010; Klauer et al., 2015). T_1 summarizes how well the model accounts for the average response frequencies across participants and is conceptually analogous to the goodness-of-fit statistic χ^2 used in non-Bayesian multinomial modeling (Klauer et al., 2015; Riefer & Batchelder, 1991). T_2 summarizes how well the model accounts for the variances and covariances of the response frequencies across participants. The posterior predictive p-value represents the comparison of the calculated test statistics obtained for observed and predicted response frequencies. A p >.05 reflects no reliable difference between observed and predicted frequencies and can be interpreted as evidence that model assumptions are in line with the data (Klauer et al., 2015; Klauer, 2010).

As model selection indices, we used the Deviance Information Criterion (DIC; Klauer et al., 2015; Spiegelhalter et al., 2002) and the Widely Applicable Information Criterion (WAIC; Vehtari et al., 2017). Models with the lowest DIC or WAIC values are interpreted to provide the best fit, and an absolute difference of 2.0 or more between models in terms of ΔDIC or ΔWAIC is generally considered as evidence in favor of one model over the other (Burnham & Anderson, 2004; Klauer et al., 2015; Spiegelhalter et al., 2002).

We used the R package TreeBUGS (Heck et al., 2018) to fit hierarchical latent-trait MPT versions of each model to the data from each measure. We used the Markov Chain Monte Carlo algorithm for three independent estimation chains with 1,000,000 iterations each, of which 250,000 were removed as a burn-in period. Every 500th iteration was retained to compute summary statistics. We report model equations and estimated parameter values in the online supplementary material. The Rubin-Gelman statistic was smaller than $\hat{R} < 1.05$ for all parameter estimates across all models, showing an acceptable convergence of estimation chains.

MPT model comparison between measures. Model fit and model selection indices are reported in Table 2 for each measure[7]. For the WIT, the PDP ($p(T_1)$ =.16; $p(T_2)$ =.26), and the traditional Quad model ($p(T_1)$ =.08; $p(T_2)$ =.09) demonstrated acceptable fit, but all other models did not demonstrate acceptable fit for at least one of the fit statistics. Selection indices suggested that the PDP provides better fit to the WIT than the traditional Quad model (ΔDIC = 70.9; ΔWAIC = 139.7) and all other models (all ΔDICs > 26.2; all ΔWAICs > 83.7).

For the FPST, only the PDP demonstrated acceptable fit ($p(T_1)$ =.05; $p(T_2)$ =.39). Selection indices suggested that the PDP provides better fit to the FPST than all other models (all ΔDICs > 8.0; all ΔWAICs > 44.4).

For the IAT, the PDP ($p(T_1)$ =.56; $p(T_2)$ =.38), PDP+G ($p(T_1)$ =.57; $p(T_2)$ =.33), and the egalitarian Quad model ($p(T_1)$ =.39; $p(T_2)$ =.33) demonstrated acceptable fit, but all other models did not demonstrate acceptable fit for at least one of the fit statistics. Selection indices suggested that the PDP provides better fit to the IAT than the PDP+G (ΔDIC = 1.9; ΔWAIC = 18.3), the

egalitarian Quad model (ΔDIC = 3.3; ΔWAIC = 18.3), and all other models (ΔDICs > 24.9; ΔWAICs > 55.9).

Insert Table 2 about here

Joint modeling across measures. In the previous section, we separately applied each model to data from each implicit measure to determine which model provided best fit to each measure. Next, we specified a joint model to all three measures simultaneously to compare correspondence in process parameters between measures.

To decide on which joint MPT model to apply to each measure, we preregistered two different selection approaches. In the first approach, we proposed to apply the best-fitting model (in terms of the lowest DIC or WAIC value) to the measure it fit best, which could result in a joint model that includes up to three different MPT models. In the second approach, we proposed to apply the same model to all three measures if it provided best fit to any of the measures. Across all models the PDP provided best model fit, so both approaches result in a joint model that includes only the PDP model (hereafter referred to as the *PDP-joint model*).

We estimated the PDP-joint model in the same way we estimated the individual models. The Rubin-Gelman statistic was smaller than $\hat{R} < 1.05$ for all parameter estimates across all models, showing an acceptable convergence of estimation chains. The PDP-joint model ($p(T_1) = .083$; $p(T_2) = .446$) fit the data well. We report parameter estimates for the PDP-only joint model in Table 3. Furthermore, we report racial bias estimates, quantified as the difference between automatic parameters for Black versus White targets, as well as parameter correlations between measures within the joint model.

Insert Table 3 about here

Parameter and racial bias estimate correlations across measures. The racial bias estimate for Black versus White males in the PDP is reflected in the difference between automatic process parameters for Black versus White faces: $\Delta A_{BW} = A_B - A_W$. We operationalize racial bias in this way analogously to how implicit bias has been quantified before (e.g., Payne, 2005; Ito et al., 2015). Positive ΔA_{BW} values (Table 3) with 95%-Bayesian Confidence Intervals (BCIs; in brackets) that do not contain zero emerged for the WIT ($\Delta A_{BW} = 0.08$ [0.05 – 0.12]), FPST ($\Delta A_{BW} = 0.07$ [0.05 – 0.10]), and IAT ($\Delta A_{BW} = 0.08$ [0.05 – 0.11]). All three measures indicate that threatening objects (i.e., guns) are reliably associated more strongly with Black than White males. Racial bias estimates correlated moderately between the WIT and FPST ($r = .38$ [.09. – .66]), but not between the IAT and WIT ($r = .16$ [-.19. – .49]) or between the IAT and FPST ($r = .11$ [-.29 – .49])[9].

C-parameters correlated strongly ($r = .52 - .65$) among measures, indicating that processes that contribute to overall task accuracy correspond highly within participants across measures. *A*-parameters correlated moderately to strongly between WIT and FPST ($r = .31 - .71$), WIT and IAT ($r = .34 - .56$), and FPST and IAT ($r = .39 - .64$). This pattern of results indicates that participants' response tendency towards "gun" over "tool" corresponded highly across measures.

Conceptually analogous *A*-parameters estimated from the WIT and FPST corresponded descriptively more strongly with one another than they did with conceptually dissimilar *A*-parameters. Specifically, A_B-parameters correlated at $r = .65$ [.45 – .81] and A_W-parameters correlated at $r = .62$ [.43 – .78], but the A_B-parameter estimated from the WIT correlated with the A_W-parameter estimated from the FPST at $r = .54$ [.35 – .72], and the A_W-parameter from the WIT correlated with the A_B-parameter estimated from the FPST at $r = .31$ [.09 – .52]. A similar, but attenuated, pattern of correlations emerged among *A*-parameters estimated from the FPST and IAT. A_B-parameters correlated at $r = .47$ [.15 – .75] and A_W-parameters correlated at $r = .50$ [.19 – .79], but the A_B-parameter estimated from the FPST correlated with the A_W-parameter estimated from the IAT at $r = .44$ [.14 – .72], and the A_W-parameter estimated from the FPST correlated with the A_B-parameter estimated from the IAT at $r = .43$ [.08 – .75]. However, a different pattern of correlations emerged among *A*-parameters estimated from the WIT and IAT. The A_W-parameter correlation $r = .56$ [.30 –

.79] across measures was stronger than both the correlation between the A_B-parameter estimated from the WIT and the A_W-parameter estimated from the IAT r = .34 [.05 – .62] and the correlation between the A_W-parameter estimated from the WIT and the A_B-parameter estimated from the IAT r = .40 [.10 – .68], but the A_B-parameter correlation r = .38, [.07 – .65] was not. Taken together, this pattern of correlations suggests a degree of theoretical precision in A-parameters, though the strength of correspondence varies across measures.

General Discussion

In the present research, we investigated correspondence among three implicit measures of racial bias: the Weapon Identification Task (WIT), the First-Person Shooter Task (FPST) and the Implicit Association Test (IAT). Previous research showing small or null correlations among these measures is confounded by differences in procedures or stimuli. In contrast, we aligned procedures and stimuli across all three measures, allowing them to vary only in structure and instructions. With measures aligned in this way, we found that participants' responses corresponded across measures, both in terms of accuracy-oriented controlled responding – replicating previous research – and also in terms of the tendency to respond "gun" versus "tool". That said, racial bias as operationalized in terms of differences between Black and White associations parameters did not correspond as consistently across measures: whereas racial bias estimates correlated moderately across the WIT and FPST, racial bias estimates corresponded weakly between the IAT and the other measures. Nevertheless, with a well-powered fully within-participants design, our findings suggest that the WIT, FPST, and IAT can correspond well with one another when procedurally aligned and analyzed with sufficient theoretical precision.

Correspondence across Implicit Racial Bias Measures

Perhaps the most important finding to emerge from the present research is that automatic process estimates correlated moderately to strongly across all measures. This finding helps to resolve a puzzle that emerged relatively early in the implicit social cognition literature, that implicit measures configured to assess the same construct corresponded poorly or not at all (Bar-Anan & Nosek, 2014; Cunningham et al., 2001; Ito et al., 2015; Olson & Fazio, 2003). Indeed, based on those findings,

implicit social cognition research was rightly criticized: If our measures do not correspond with one another, do we even know what we are measuring? However, by aligning measures across procedures and stimuli and using MPT modeling to disentangle the joint contributions of multiple processes, we showed that measures configured to assess the same construct correspond well with one another in automatic process estimates for the same target group.

The PDP provided best fit to all three implicit measures examined in the present research, which begs the question: What cognitive process(es) does the automatic parameter of the PDP reflect? The most straightforward interpretation of the automatic parameter is that it reflects participants' preference to respond with gun in comparison to tool. This preference can reflect a simple hand-side preference, other dispositional characteristics like a threat-related attention bias driven by anxiety (Bar-Haim et al., 2007), or a general threat superiority effect for gun targets (Rivera-Rodriguez et al., 2021; Subra et al., 2018). Additionally, automatic parameters provide insight into racial biases when we compare parameters estimated from trials that include Black versus White faces. Across all measures, participants demonstrated stronger associations between "Black" and "gun" than between "White" and "gun", illustrating the cultural stereotype of Black men as dangerous – which, in turn, provides a degree of validity evidence that these measures can assess their intended construct.

Though automatic parameters correlated moderately-to-strongly across all tasks, racial bias operationalized as the difference between Black and White automatic parameters did not align in the same way. Specifically, this operationalization of racial bias correlated moderately between the WIT and FPST, but not between the IAT and the other two measures. These findings dovetail with previous work demonstrating correspondence between the WIT and FPST (Payne & Correll, 2020; Ito et al., 2015). In contrast, lack of correspondence between the IAT and the other two measures may reflect heightened category salience and direct contrasting of race categories as a function of the structure of the IAT. In all three tasks, participants view pictures of Black and White males, but never need to attend to race – or to the faces at all – to make a correct response on the WIT (i.e., "gun", "tool") or the FPST (i.e., "shoot", "don't shoot"). Only the IAT, because of its dual-categorization structure, requires participants to attend to race to make a correct response to half of

trials. Thus, our findings would seem to align with other researchers who conclude that the IAT assesses associations based on racial categories (De Houwer, 2001), but the WIT and FPST assess associations based on racial exemplars (Olson & Fazio; Livingston & Brewer, 2002). That said, difference score-based racial bias estimates may instead correlate less consistently across measures than associations parameters for statistical reasons: Difference scores reflect compounded variance, which may attenuate correlations (Gardner & Neufeld, 1987). Nevertheless, future research is necessary to better understand factors that moderate correspondence among processes that contribute to racially-biased responses on implicit measures.

Replicating previous research, controlled process estimates overall correlated strongly across all measures. This pattern of results indicates that participants' cognitive control abilities generalize across tasks. These findings dovetail with other work showing that the WIT and FPST are related to higher cognitive executive functions like inhibition (Ito et al., 2015; Payne, 2005), and the IAT is related to updating and task switching (Ito et al., 2015; Klauer et al., 2010). Similar findings of correspondence have also emerged in research using other accuracy-based measures of conflict resolution (Draheim et al., 2021). Hence, the present research joins a body of literature connecting implicit measures with a broad constellation of executive functions and other higher cognitive abilities.

Summary statistics of implicit bias

Racial bias effects emerged across all three measures in terms of accuracy-based summary statistics, such that guns are associated more strongly with Black than White male faces. However, these accuracy-based summary statistics only corresponded weakly between the WIT and FPST, and not between the IAT and either the WIT or the FPST. In contrast, MPT modeling revealed correspondence among processes across measures, such that both control and automatic parameters correspond well to conceptually-analogous parameters across measures. In addition, racial bias estimates operationalized in terms of differences between model parameters correlated moderately between the WIT and FPST, demonstrating that MPT modeling can reveal correspondence between measures that summary statistics may obscure. Taken together, the present research highlights one

way in which summary statistics of implicit bias can belie process-level relationships, and at the same time demonstrates the value of the theoretical precision provided by MPT models.

Multinomial process tree models

Not only does the present research help to resolve an apparent puzzle identified by previous research, but it also points the way forward for future research using implicit measures. We relied on MPT modeling to estimate the contributions of qualitatively distinct cognitive processes to responses on different implicit measures. Because MPT models reflect theoretical assumptions instantiated in equation form, the degree to which each MPT fits data provides insight into the cognitive processes that underlie responses on each implicit measure.

The PDP provided unambiguously best fit to the WIT, which is perhaps unsurprising because the PDP has been extensively used with the WIT in previous research (e.g., Bishara & Payne, 2009; Ito et al., 2015; Klauer et al., 2015; Klauer & Voss, 2008; Lambert et al., 2003; Payne, 2001; Payne et al., 2002). The traditional Quad model also provided acceptable fit to the WIT, but model selection indices favored the PDP over it. The PDP also provided unambiguously best fit to the FPST, which is also perhaps unsurprising. Though the FPST is typically analyzed using signal detection theory (SDT) modeling (Correll et al., 2002; Mekawi & Bresin, 2015), the PDP and SDT are mathematically very similar (Payne & Correll, 2020). The PDP also provided best fit to the IAT, though the egalitarian Quad model and PDP+G model also provided acceptable fit. Taken together, the present research underscores the value of the PDP to disentangle the contributions of distinct processes to three widely-used implicit measures. At the same time, our findings support the utility of the Quad model (in the context of the WIT and IAT) and PDP+G model (in the context of the IAT) for researchers who seek to investigate processes not accounted for by parameters of the PDP.

Just as the present research identified MPT models that can be validly applied to the three implicit measures we investigated, so too did it identify models that provide poor fit to these measures. Specifically, the Stroop, Stroop+G, and SMT models all did not provide acceptable fit to any of the three measures. Notably, each of these models reflect the assumption that the relatively more automatic process dominates responses – in contrast to the PDP, PDP+G, and Quad models

that assume relatively more controlled processes dominate responses. Thus, one straightforward takeaway from the present research is that controlled processes will generally drive responses on the WIT, FPST, and IAT if given the opportunity. With that said, we recognize that any single sample is insufficient to comprehensively declare the Stroop, Stroop+G, and SMT models as invalid for use with these three implicit measures. Nevertheless, our findings highlight the dominance of controlled processes in these measures, but future research may identify contexts, participant samples, or other moderators under which automatic processes are relatively more influential on responses.

Limitations

Despite the strengths of the present research, it is also limited in some ways. For example, to test our primary research question about correspondence among implicit measures, we aligned all three measures in terms of stimuli and procedures. Such alignment positioned us to make apples-to-apples comparisons among measures, but this comparability came at the cost of modifying some measures from their traditional form. Perhaps the biggest modification was the inclusion of neutral face shapes, which was necessary to provide sufficient degrees of freedom for MPT modeling. The inclusion of neutral face shapes would not seem to deviate significantly from the traditional forms of the WIT or FPST, given that the instructions remained unchanged in our modified versions of these measures. However, the inclusion of the neutral face shapes changed the structure of the IAT because an additional set of blocks was added to accommodate this third category (i.e., Black/White, Black/neutral, White/neutral). That said, this expanded IAT format closely aligns with an existing version of the IAT: the multi-category IAT (Axt et al., 2014). Moreover, the expected bias effects (i.e., indicating stronger associations between threat and Black versus White) emerged across all three measures, which suggest that these changes did not significantly alter the measures. Nevertheless, previous research has demonstrated that task procedures (such as number of trials) affect the extent to which different cognitive processes influence responses on implicit measures (Calanchini et al., 2021), so future research should continue to investigate the role of stimulus and structural effects in implicit bias.

Conclusion

The present research makes two key contributions to the implicit social cognition literature. The first contribution is the demonstration that three commonly-used implicit measures can correspond well when procedurally aligned and analyzed with sufficient theoretical precision. The second contribution is our finding that the PDP specifically, and control-dominant MPTs more generally, are the most suitable models for these measures. Taken together, this work offers a useful template for future researchers who seek to incorporate multiple operationalizations of racial bias in their work and gain theoretically-precise insight into the processes that contribute to responses on measures of implicit social cognition.

References

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661. https://doi.org/10.1037/0022-3514.91.4.652

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The Rules of Implicit Evaluation by Race, Religion, and Age. *Psychological Science, 25*(9), 1804–1815. https://doi.org/10.1177/0956797614543801

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133*(1), 1–24. https://doi.org/10.1037/0033-2909.133.1.1

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior research methods*, *46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Bishara, A. J., & Payne, B. K. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology*, *45*(3), 524–534. https://doi.org/10.1016/j.jesp.2008.11.002

Burke, C. T. (2015). Process dissociation models in racial bias research: Updating the analytic method and integrating with signal detection approaches. *Group Processes & Intergroup Relations*, *18*(3), 402–434. https://doi.org/10.1177/1368430214567763

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Calanchini, J. (2020). How multinomial processing trees have advanced, and can continue to advance, research using implicit measures. *Social Cognition*, *38*(Supplement), s165-s186. https://doi.org/10.1521/soco.2020.38.supp.s165

Calanchini, J., Meissner. F., & Klauer, K. C. (2021). The role of recoding in implicit social cognition: Investigating the scope and interpretation of the ReAL model for the implicit association test. *PLOS ONE, 16(4)*, e0250068. https://doi.org/10.1371/journal.pone.0250068

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Conrey, 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*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The Police Officer's Dilemma: A Decade of Research on Racial Bias in the Decision to Shoot: The Police Officer's Dilemma. *Social and Personality Psychology Compass*, *8*(5), 201–213. https://doi.org/10.1111/spc3.12099

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023. https://doi.org/10.1037/0022-3514.92.6.1006

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology*, *108*(2), 219–233. https://doi.org/10.1037/pspa0000015

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit Attitude Measures: Consistency, Stability, and Convergent Validity. *Psychological Science, 12*(2), 163–170. https://doi.org/10.1111/1467-9280.00328

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*(6), 443–451. https://doi.org/10.1006/jesp.2000.1464

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General, 150*(2), 242–275. https://doi.org/10.1037/xge0000783

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial Processing Tree Models: A Review of the Literature. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Foroni, F., & Bel-Bahar, T. (2010). Picture-IAT versus Word-IAT: level of stimulus representation influences on the IAT. *European Journal of Social Psychology*, *40*(2), 321–337. https://doi.org/10.1002/ejsp.626

Frenken, M., Hemmerich, W., Izydorczyk, D., Scharf, S., & Imhoff, R. (2022). Cognitive processes behind the shooter bias: Dissecting response bias, motor preparation and information accumulation. *Journal of Experimental Social Psychology*, *98*, 104230. https://doi.org/10.1016/j.jesp.2021.104230

Gardner, R. C., & Neufeld, R. W. J. (1987). Use of the Simple Change Score in Correlational Analyses'. *Educational and Psychological Measurement, 47*(4), 849–864. https://doi.org/10.1177/0013164487474001

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects? *Cognition and Emotion*, *24*(6), 1008–1025. https://doi.org/10.1080/02699930903112712

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, *44*(1), 164–172. https://doi.org/10.1016/j.jesp.2007.01.002

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual review of psychology*, *71*, 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

Greenwald, 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*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

Huntsinger, J. R., Sinclair, S., & Clore, G. L. (2009). Affective regulation of implicitly measured stereotypes and attitudes: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *45*(3), 560–566. https://doi.org/10.1016/j.jesp.2009.01.007

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology, 27*, 116–159. https://doi.org/10.1080/10463283.2016.1212966

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Klauer, K. C. (2010). Hierarchical Multinomial Processing Tree Models: A Latent-Trait Approach. *Psychometrika*, *75*(1), 70–98. http://dx.doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221. https://doi.org/10.1037/xge0000044

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *Quarterly Journal of Experimental Psychology*, *63*(3), 595–619. https://doi.org/10.1080/17470210903076826

Klauer, K. C., & Voss, A. (2008). Effects of Race on Responses and Response Latencies in the Weapon Identification Task: A Test of Six Models. *Personality and Social Psychology Bulletin*, *34*(8), 1124–1140. https://doi.org/10.1177/0146167208318603

Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, *103*(2), 205–224. https://doi.org/10.1037/a0028764

Lambert, A. J., Payne, B. K., Jacoby, L. L., Shaffer, L. M., Chasteen, A. L., & Khan, S. R. (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology*, *84*(2), 277–295. https://doi.org/10.1037/0022-3514.84.2.277

Laukenmann, R., Erdfelder, E., Heck, D. W., & Moshagen, M. (2023). Cognitive processes underlying the weapon identification task: A comparison of models accounting for both response frequencies and response times. *Social Cognition*, *41*(2), 137–164. https://doi.org/10.1521/soco.2023.41.2.137

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance, 20*(2), 219–234. https://doi.org/10.1037/0096-1523.20.2.219

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*(1), 5–18. https://doi.org/10.1037/0022-3514.82.1.5

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*(1), 45–69. https://doi.org/10.1037/a0030734

Mekawi, Y., & Bresin, K. (2015). Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology*, *61*, 120–130. https://doi.org/10.1016/j.jesp.2015.08.002

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science*, *14*(6), 636–639. https://doi.org/10.1046/j.0956-7976.2003.psci_1477.x

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192. https://doi.org/10.1037/0022-3514.81.2.181

Payne, B. K. (2005). Conceptualizing Control in Social Cognition: How Executive Functioning Modulates the Expression of Automatic Stereotyping. *Journal of Personality and Social Psychology*, *89*(4), 488–503. https://doi.org/10.1037/0022-3514.89.4.488

Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. In *Advances in experimental social psychology* (Vol. 62, pp. 1-50). Academic Press. https://doi.org/10.1016/bs.aesp.2020.04.001

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology*, *38*(4), 384–396. https://doi.org/10.1016/S0022-1031(02)00006-9

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology*, *41*(1), 36–47. https://doi.org/10.1016/j.jesp.2004.05.001

Phills, C. E., Kawakami, K., Tabi, E., Nadolny, D., & Inzlicht, M. (2011). Mind the gap: Increasing associations between the self and blacks with approach behaviors. *Journal of Personality and Social Psychology, 100*(2), 197–210. https://doi.org/10.1037/a0022159

Plant, E. A., Peruche, B. M., & Butz, D. A. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology*, *41*(2), 141–156. https://doi.org/10.1016/j.jesp.2004.07.004

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*(3), 318–339. https://doi.org/10.1037/0033-295X.95.3.318

Riefer, D. M., & Batchelder, W. H. (1991). Statistical Inference for Multinomial Processing Tree Models. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical Psychology: Current Developments* (pp. 313–335). Springer. https://doi.org/10.1007/978-1-4613-9728-1_18

Rivera-Rodriguez, A., Sherwood, M., Fitzroy, A. B., Sanders, L. D., & Dasgupta, N. (2021). Anger, race, and the neurocognition of threat: attention, inhibition, and error processing during a weapon identification task. *Cognitive research: principles and implications*, *6*, 1–27. https://doi.org/10.1186/s41235-021-00342-w

Rivers, A. M. (2017). The Weapons Identification Task: Recommendations for adequately powered research. *PLOS ONE*, *12*(6), e0177857. https://doi.org/10.1371/journal.pone.0177857

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Schimmack, U. (2021). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science, 16*(2), 396–414. https://doi.org/10.1177/1745691619863798

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, *56*(4), 283–294. https://doi.org/10.1027/1618-3169.56.4.283

Stein, T., Ciorli, T., & Otten, M. (2023). Guns Are Not Faster to Enter Awareness After Seeing a Black Face: Absence of Race-Priming in a Gun/Tool Task During Continuous Flash Suppression. *Personality and Social Psychology Bulletin, 49*(3), 405–414. https://doi.org/10.1177/01461672211067068

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Subra, B., Muller, D., Fourgassie, L., Chauvin, A., & Alexopoulos, T. (2018). Of guns and snakes: testing a modern threat superiority effect. *Cognition and emotion*, *32*(1), 81–91. https://doi.org/10.1080/02699931.2017.1284044

Thiem, K. C., Neel, R., Simpson, A. J., & Todd, A. R. (2019). Are Black Women and Girls Associated With Danger? Implicit Racial Bias at the Intersection of Target Age and Gender. *Personality and Social Psychology Bulletin, 45*(10), 1427–1439. https://doi.org/10.1177/0146167219829182

Todd, A. R., Johnson, D. J., Lassetter, B., Neel, R., Simpson, A. J., & Cesario, J. (2021). Category salience and racial bias in weapon identification: A diffusion modeling approach. *Journal of Personality and Social Psychology, 120*(3), 672–693. https://doi.org/10.1037/pspi0000279

Unkelbach, C., Forgas, J. P., & Denson, T. F. (2008). The turban effect: The influence of Muslim headgear and induced affect on aggressive responses in the shooter bias paradigm. *Journal of Experimental Social Psychology*, *44*(5), 1409–1413. https://doi.org/10.1016/j.jesp.2008.04.003

Unkelbach, C., Goldenberg, L., Müller, N., Sobbe, G. & Spannaus, N. (2009). A shooter bias in Germany against people wearing Muslims headgear. *Revue internationale de psychologie sociale*, 22, 181–201. https://www.cairn.info/revue--2009-3-page-181.htm.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Volpert-Esmond, H. I., Scherer, L. D., & Bartholow, B. D. (2020). Dissociating automatic associations: Comparing two implicit measurements of race bias. *European Journal of Social Psychology*, *50*(4), 876–888. https://doi.org/10.1002/ejsp.2655

Footnotes

[1] We use the term "implicit" to mean "indirect". Thus, the term "implicit measure" refers to a family of measurement instruments which assess mental contents without directly asking participants for that information.

[2] We use the terms "implicit bias" to refer to behavioral responses on implicit measures, and the term "associations" to refer to the underlying mental construct assessed by implicit measures. We make no strong assumptions or claims about the representational nature of the constructs assessed by implicit measures.

[3] A recent version of the FPST also allows for a dynamic presentation of target stimuli by showing a video clip of a person reaching into their pocket and pulling out the target object, resulting in a sequential presentation of the person first and the target object second (Frenken et al., 2022).

[4] In the present research, we defined the *A*-parameter as an Automatic process influencing responses towards "gun" on all trial types. This specification contrasts with the specification proposed by Bishara and Payne (2009), in which the *A*-parameter is defined as influencing responses towards "gun" for Black faces but towards "tool" for White faces.

[5] In a first wave of preliminary data collection ($N = 42$), we imposed a response time limit of 500 ms after target onset (with a pattern mask shown for 300 ms) which resulted in a substantially higher error rate for the WIT ($M = 33.9\%$) compared to the FPST ($M = 13.8\%$) and the IAT ($M = 12.6\%$). To improve comparability between measures and their process models, we aim to have similar error rates for participants across measures and thus adjusted the response time limit to 700ms for the WIT. These preliminary data were not included in further analysis.

[6] We report means, standard deviations and further statistical analysis of accuracy and response time data by target and race for each measure in the online supplementary material.

[7] We constrained process parameters reflecting controlled responding and guessing across race conditions but allowed parameters reflecting automatic influences to vary between race conditions. These constraints allow for comparison across models with degrees of freedom > 0.

[8] In an exploratory fashion, we estimated alternative joint models with applying the egalitarian or traditional Quad model to the IAT and the PDP to the WIT and FPST. We report both models in the online supplementary material.

[9] Note that correlations based on difference scores should be interpreted with caution (Gardner & Neufeld, 1987).

Tables

Table 1

Procedural details of the WIT, FPST and IAT in their originally published versions.

	Weapon Identification Task (WIT; Payne, 2001)	First-person Shooter Task (FPST; (Correll et al. 2002)	Implicit Association Test (IAT; Greenwald et al., 1998; Study 3)
response time limit	none (Study 1); 500 ms (Study 2)	850ms (Study 1 & 3); 630ms (Study 2)	none
number of practice trials	48 trials	no trials	150 trials
number of experimental trials	192 trials	80 trials	200 trials
stimulus material: target group / category	cropped face images of Black and White males as primes	full body image of Black and White males	first names judged to be more likely to belong to Black or White persons
stimulus material: target object / attribute	target images of weapons and tools	target weapon object or an innocuous object (e.g., a cell phone, a wallet) held by Black and White males	unpleasant and pleasant words as attributes
presentation order of stimulus material	sequential: prime face followed by target objects presented in the center of the screen	concurrent: person holding target object presented at a random position on screen	serial: target category or attribute stimuli presented in the center of the screen in a random order between target and attribute stimuli
instruction	identifying target object while ignoring face prime	"shoot" armed person, "don't shoot" unarmed person	correctly categorize the target category and attribute stimuli
analysis	correct response latencies, error rates, process dissociation procedure	correct response latencies, error rates, signal detection modeling	correct response latencies

Table 2

MPT-model fit and selection indices for the three implicit measures.

	WIT				*FPST*				*IAT*			
MPT model	$p(T_1)$	$p(T_2)$	*DIC*	*WAIC*	$p(T_1)$	$p(T_2)$	*DIC*	*WAIC*	$p(T_1)$	$p(T_2)$	*DIC*	*WAIC*
PDP	.164	.260	**10003.4**	**9898.1**	.053	.385	**9399.1**	**9294.3**	.558	.385	**9159.8**	**9073.3**
PDP+G	.010	.210	10029.6	9985.2	.042	.270	9407.1	9338.7	.570	.334	9161.7	9091.6
Stroop	<.001	<.001	10503.1	10605.6	<.001	<.001	9808.4	9872.0	<.001	<.001	9403.4	9439.8
Stroop+G	<.001	.089	10112.3	10087.9	<.001	.109	9427.8	9367.8	<.001	.032	9215.2	9172.7
traditional Quad	.079	.086	10074.3	10037.8	.023	.154	9410.6	9344.1	<.001	.068	9184.7	9129.2
egalitarian Quad	.016	.202	10029.9	9981.8	.033	.262	9408.8	9340.3	.391	.330	9163.1	9091.6
SMT-model	.003	.175	10077.3	10005.4	<.001	.112	9431.5	9368.9	<.001	.036	9218.3	9175.1

Note. $p(T_1)$ and $p(T_2)$ correspond to the p-values associated with the T_1 and T_2 indices, respectively, and values <.05 indicate poor fit. The lowest DIC and WAIC indicates the favored model and is printed in bold.

Table 3

Mean parameter estimates and intercorrelations for PDP-only joint model.

Parameter	*Mean*	*95%-BCI*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
WIT																
1. C	.748	[.729 – .767]														
2. A_B	.542	[.519 – .564]	-.13													
3. A_W	.459	[.436 – .483]	-.06	.15												
4. A_N	.460	[.438 – .482]	-.04	**.43**	**.51**											
5. ΔA_{BW}	0.083	[0.051 – 0.115]	-.03	**.61**	**-.69**	-.16										
FPST																
6. C	.797	[.784 – .810]	**.56**	-.18	-.12	-.05	-.03									
7. A_B	.507	[.484 – .529]	-.13	**.65**	**.31**	**.56**	**.22**	-.09								
8. A_W	.433	[.411 – .456]	-.10	**.54**	**.62**	**.71**	-.10	**-.21**	**.64**							
9. A_N	.438	[.414 – .461]	**-.18**	**.46**	**.51**	**.70**	-.07	-.08	**.65**	**.76**						
10. ΔA_{BW}	0.074	[0.045 – 0.103]	-.03	.10	**-.37**	-.19	**.38**	.14	**.39**	**-.44**	-.15					
IAT																
11. C	.821	[.810 – .831]	**.52**	-.08	.08	.02	-.12	**.65**	-.10	-.08	.02	-.02				
12. A_B	.533	[.512 – .554]	.18	**.38**	**.40**	**.46**	-.04	.24	**.47**	**.43**	**.39**	.03	.20			
13. A_W	.457	[.434 – .478]	-.13	**.34**	**.56**	**.47**	-.20	-.16	**.44**	**.50**	**.43**	-.09	-.02	.36		
14. A_N	.405	[.382 – .427]	-.10	**.46**	**.51**	**.56**	-.07	**-.23**	**.56**	**.64**	**.63**	-.10	.05	.32	**.55**	
15. ΔA_{BW}	0.076	[0.047 – 0.106]	**.26**	-.02	-.22	-.08	.16	**.34**	-.05	-.14	-.10	.11	.18	.44	**-.65**	-.26

Note. Correlations are calculated on the probit scale. Correlations in bold indicate that the 95%-BCI does not contain zero. BCI = Bayesian Credibility Interval. $\Delta A_{BW} = A_B - A_W$.

Figures

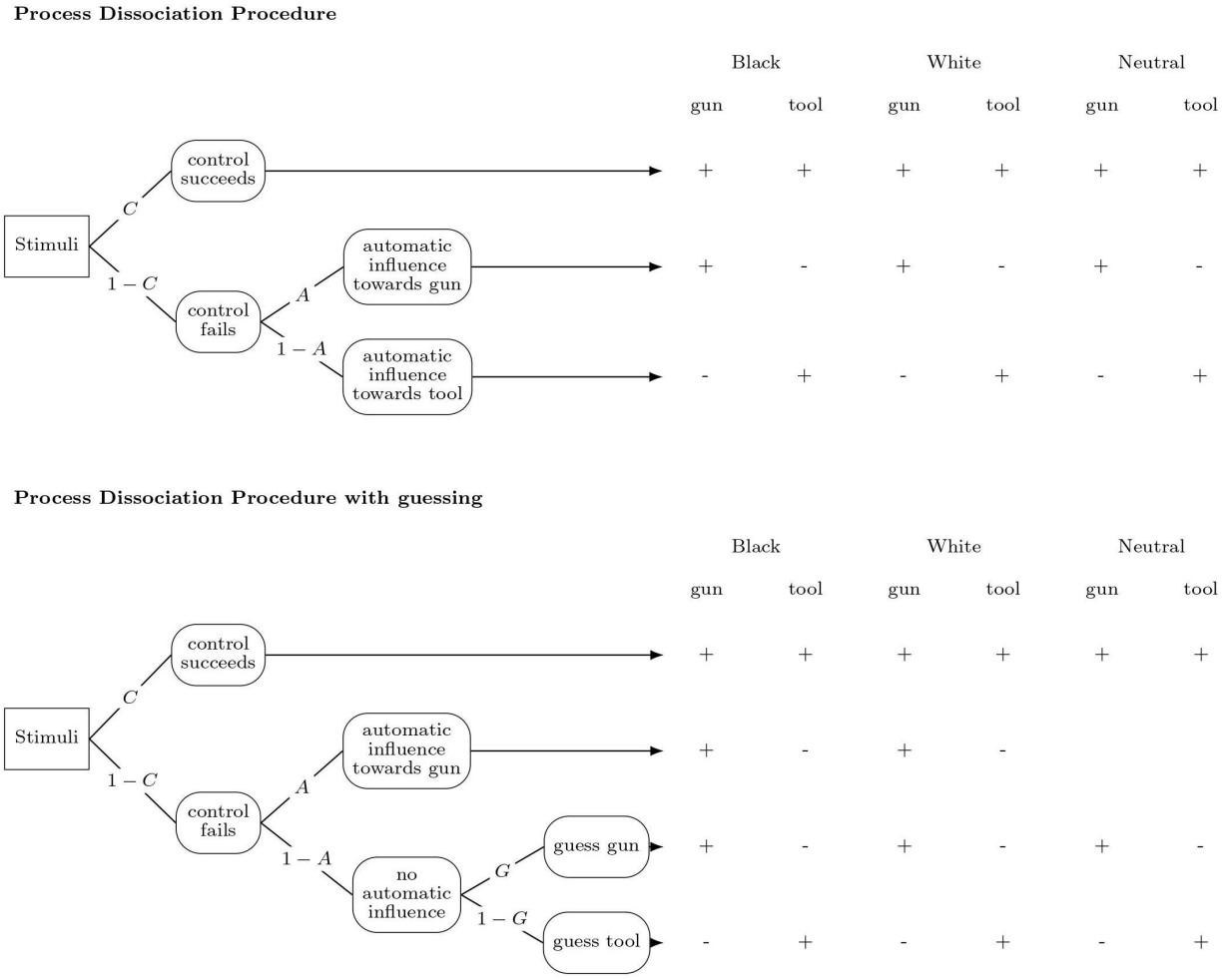


Figure 1. Multinomial processing trees of the process dissociation procedure (PDP; upper panel) and of the PDP with guessing (lower panel). Branches lead to correct (+) and incorrect (–) responses. C = Controlled process, A = Automatic process, and G = Guessing.

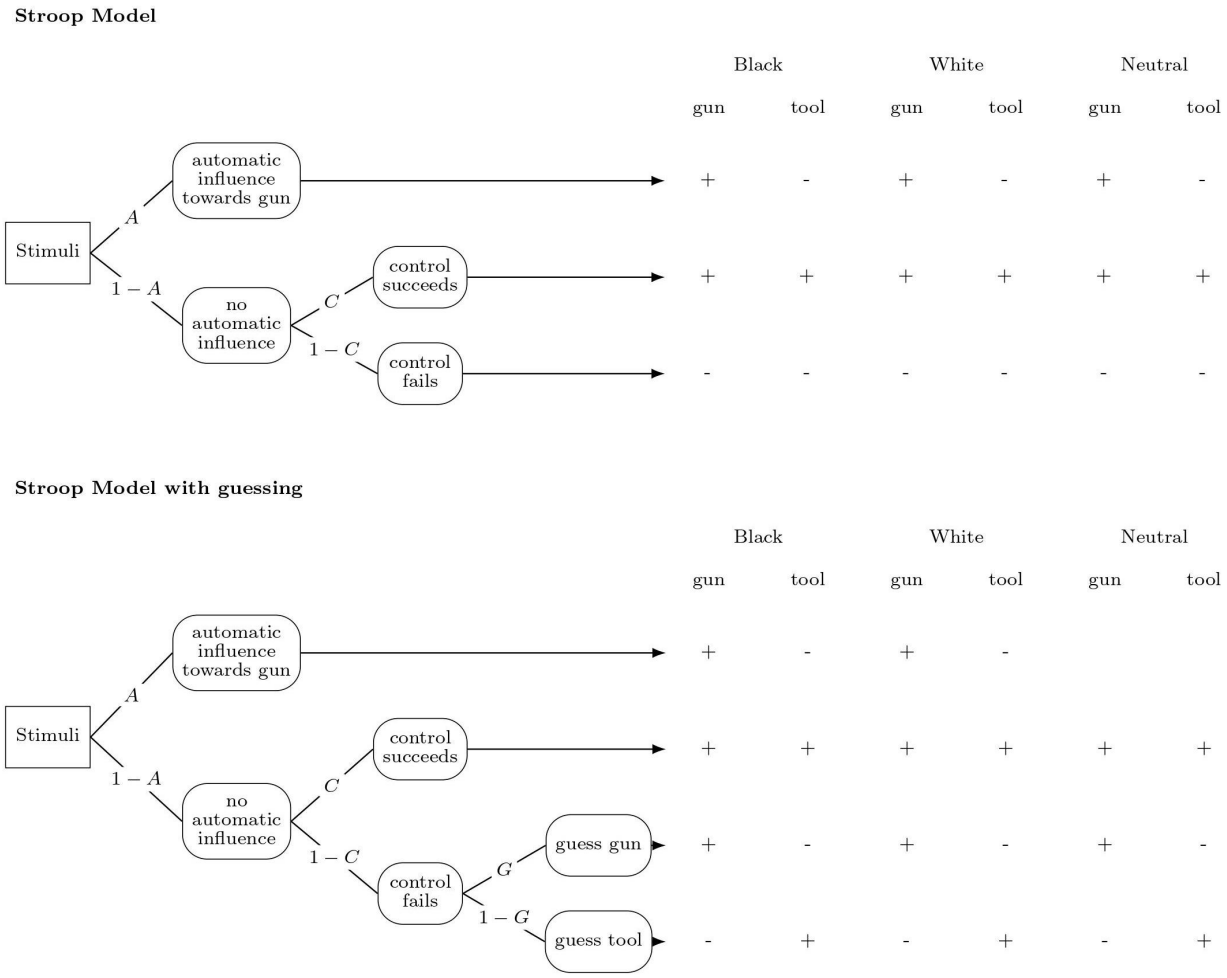


Figure 2. Multinomial processing trees of the Stoop Model (upper panel) and of the Stroop Model with guessing (lower panel). Branches lead to correct (+) and incorrect (–) responses. C = Controlled process, A = Automatic process, and G = Guessing.

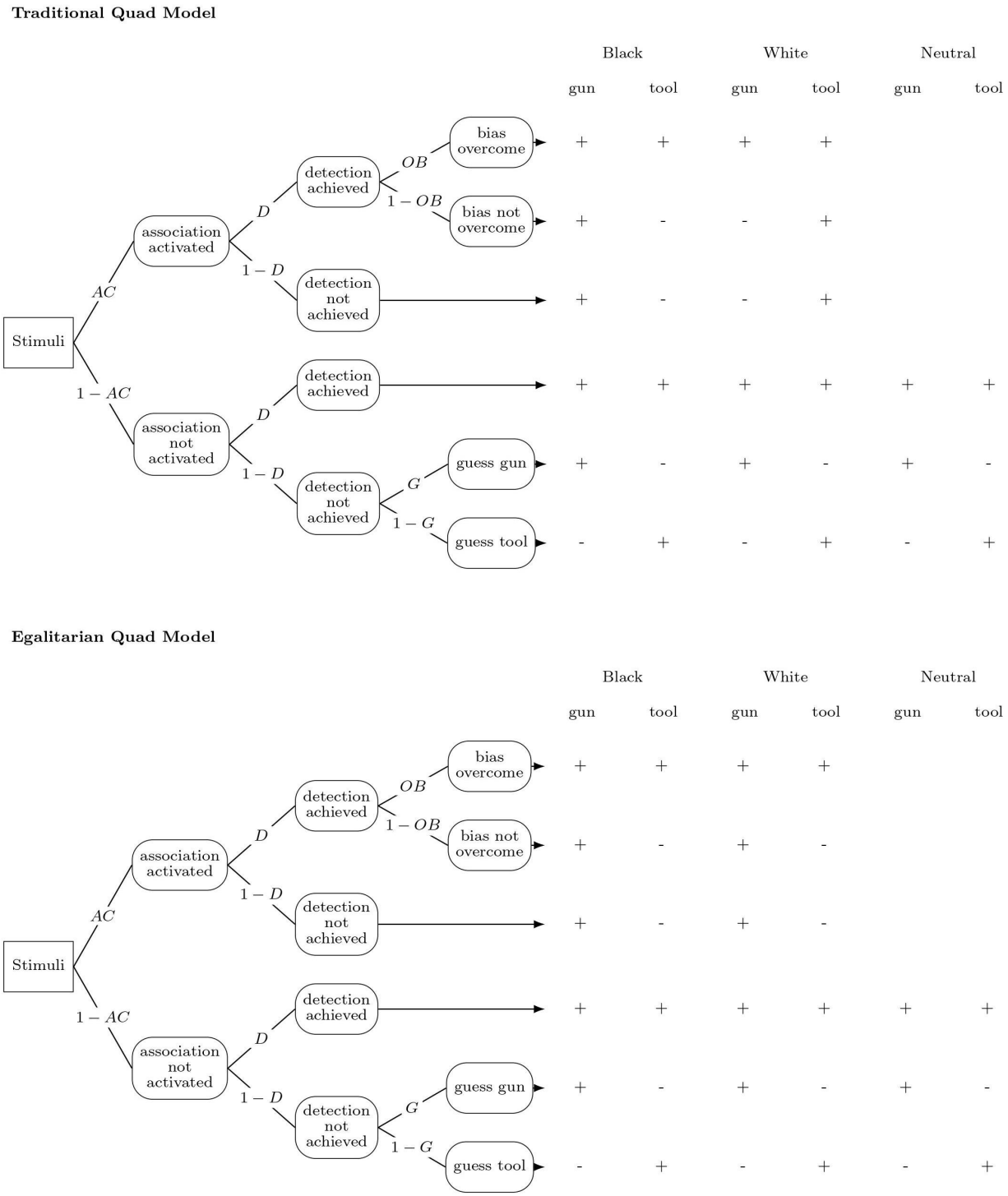


Figure 3. Multinomial Processing Tree of the traditional Quad model (upper panel) and the egalitarian Quad model (lower panel). Branches lead to correct (+) and incorrect (–) responses. $D =$ Discrimination, $AC =$ Association Activation, $OB =$ Overcoming Bias, and $G =$ Guessing.

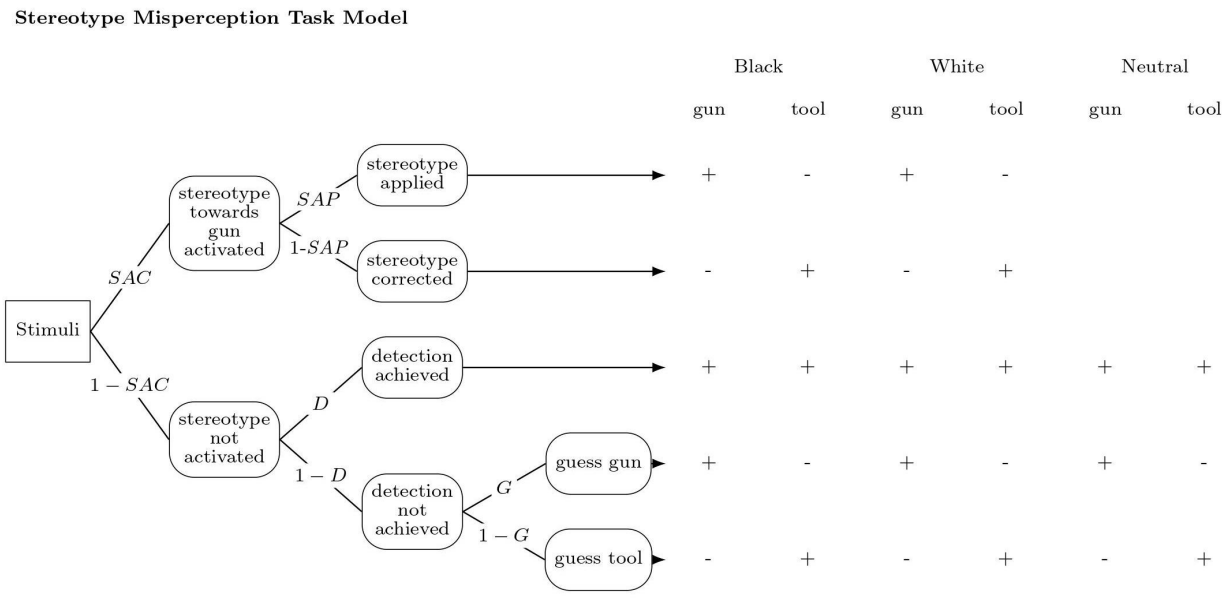


Figure 4. Multinomial processing tree of the Stereotype Misperception Task Model (SMT Model). Branches lead to correct (+) and incorrect (–) responses. D = Detection, SAC = Association Activation, SAP = Association Application, and G = Guessing.

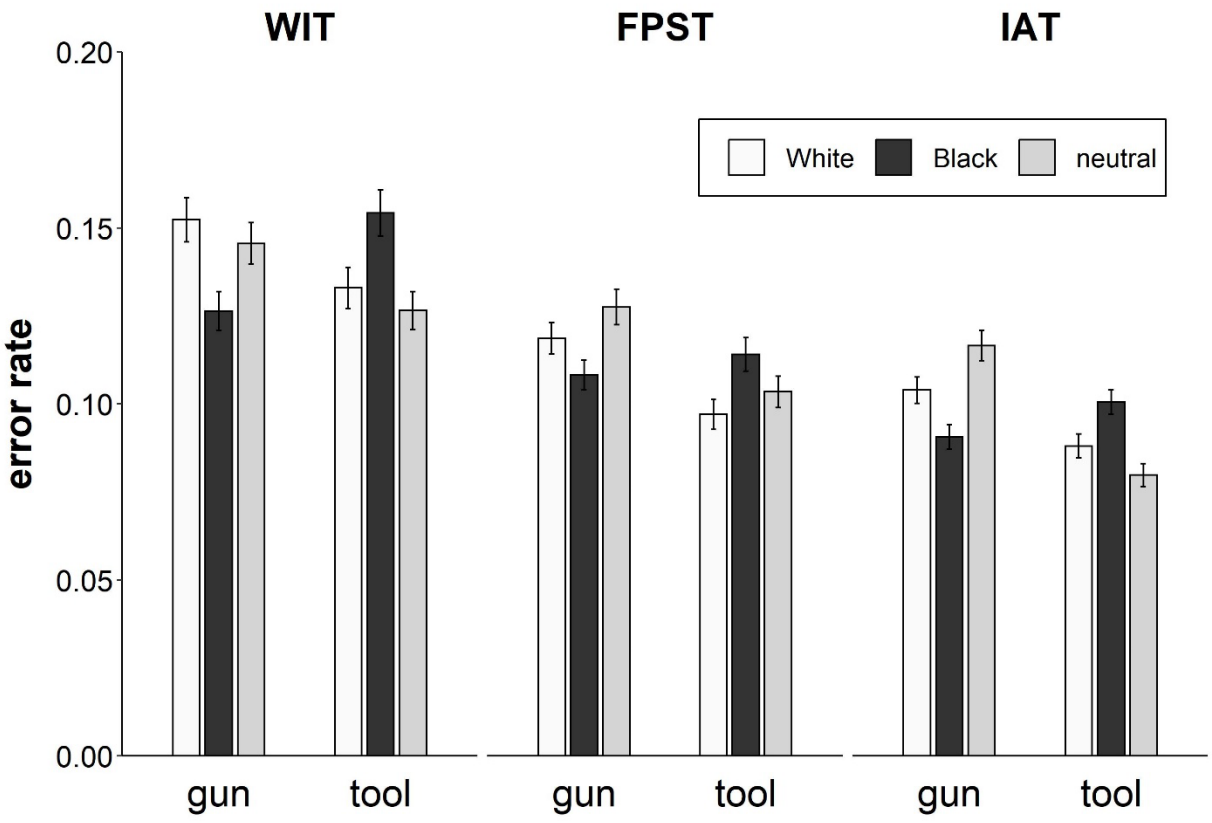


Figure 5. Error rates of the three implicit measures by race and target. WIT = Weapon Identification Task, FPST = First-Person Shooter Task, IAT = Implicit Association Task. Error bars represent one standard error.

This dissertation was supported by the Research Training Group "Statistical Modeling in Psychology", funded by the Deutsche Forschungsgemeinschaft (GRK 2277), by a Doctoral Research Fellowship granted by the Foundation of German Business (Stiftung der Deutschen Wirtschaft, sdw gGmbH), funded by the German Federal Ministry of Education and Research, and by a one-year research grant for doctoral candidates funded by the German Academic Exchange Service (DAAD).