Supplementary Material

Post-task Questions

Following both tasks in Study 1, participants responded to the following questions using scales anchored at 1 (*not at all*) and 7 (*extremely*): "How frustrated were you by your errors?"; "How anxious did your errors make you?"; "How unpleasant were your errors?"; "In general, how attentive were you during the task?"; and "In general, how hard did you try during the task?" An error-related distress composite, created by averaging the first three of these items, showed acceptable internal consistency (α = .79 following the WIT; α = .71 following the APT). Additionally, responses to the attention and effort items were strongly correlated ($r = .71$, $p < .001$), and all participants reported satisfactory levels of attention ($M = 5.68$ in WIT; $M = 5.61$ in APT) and effort ($M = 5.96$ in WIT; $M = 5.93$ in APT).

Using a Difference Score Approach for PDP-A Estimates

In the main text we reported analyses of task-wide PDP-A estimates using residual scores to control for prime race effects (i.e., regressing White prime *A* estimates from Black-prime *A* estimates). Here, we report parallel analyses using the difference between Black-prime and White-prime *A* estimates (i.e., Black – White). The mean of Black-prime *C* and White-prime *C* estimates were still used as the task-wide PDP-C estimate.

Comparing PDP estimates across tasks. In the main text, PDP-A estimates calculated using residual scores were weakly correlated in Study 1, $\beta = .23$, $p = .026$, and in Study 2, $\beta = .26$, $p < .001$, showing smaller associations across tasks than PDP-C estimates. A similar comparison was done using PDP-A estimates calculated as a

difference score. Difference score PDP-A estimates were not correlated in Study 1, β = .04, p = .649, but significantly correlated in Study 2, β = .20, p = .006. More importantly, a multiple regression modeling the Task x PDP estimate interaction using the difference score PDP-A estimates showed a significant interaction in both Study 1, β = .57, p < .001, and Study 2, β = .45, p < .001 (Figure S1), replicating the pattern in the main text; namely, that PDP-A scores were less correlated than PDP-C scores across tasks.

As a whole, these results mirror the pattern of results found when the residual PDP-A estimate was used, as reported in the main text. Despite similarities, we felt using a residual was more theoretically appropriate and therefore highlighted those analyses in the main text.

Comparison of Study 1 and Study 2

The degree to which accuracy bias, PDP-Auto and PDP-Control corresponded between the WIT and APT were compared across the two studies. These models included a predictor for Study (Study 1 vs. Study 2) and an interaction term involving Study and the (standardized) outcome from the WIT predicting the (standardized) outcome from the APT. A significant interaction indicates that the slope describing the association of the outcome measure in question from the two tasks differed across the two studies.

The model examining accuracy bias across studies showed a marginal Study x WIT interaction, β = .22, t = 1.84, p = .067, indicating that the slopes describing the association of WIT and APT accuracy bias estimates were not significantly different across Study 1 (β = .19) and Study 2 (β = .41) using traditional significance levels. The model examining PDP-Auto estimates across studies showed a significant Study x WIT

interaction, $\beta = .26$, $t = 2.06$, $p = .040$, indicating that the slopes describing the association of PDP-Auto derived from the WIT and APT differed significantly across Study 1 ($\beta = .06$) and Study 2 ($\beta = .32$). Finally, the model examining PDP-Control estimates across studies showed a nonsignificant Study x WIT interaction, $\beta = .09$, $t = 0.92$, $p = .368$, indicating that the slopes describing the association of PDP-Control estimates derived from the two tasks did not differ across studies.

Reliability of PDP Estimates

Comparing the magnitude of correlations across different sets of variables relies on the assumption that those variables are measured with comparable reliability. We tested that assumption in the current data by calculating split-half reliability estimates for both PDP-A and PDP-C in both tasks. These calculations showed that although PDP-C estimates (WIT: $r = .69$, 95% [.56 - .78]; APT: $r = .71$, 95% [.59 - .80]) had somewhat higher reliability than PDP-A estimates (WIT: $r = .58$, 95% [.43 - .70]; APT: $r = .54$, 95% [.38 - .67]), their confidence intervals overlapped, indicating that those apparent differences are not particularly meaningful. Moreover, estimates of PDP components across tasks were highly similar, indicating similar reliabilities across the tasks.

Study 2 data were consistent with the Study 1 data, PDP-C estimates (WIT: $r = .73$, 95% [.65 - .79]; APT: $r = .52$, 95% [.41 - .61]) had somewhat higher reliability than PDP-A estimates (WIT: $r = .50$, 95% [.39 - .60]; APT: $r = .50$, 95% [.39 - .60]). Unlike in Study 1, the reliability of the PDP-C estimate in the WIT was higher than in the APT, and higher than PDP-A in both tasks.

Internal and External Influences on Implicit Bias

The current studies additionally explored the influence of internal and external influences on both automatic and controlled processing in these tasks. Early theorizing suggested that measurement of implicit bias using speeded response tasks is resistant to voluntary control and the influence of social norms exerted by external factors (Greenwald, McGhee, & Schwartz, 1998). However, some research has shown that factors external to the participant can moderate expression of bias during such tasks, but findings have been mixed. For example, *decreased* implicit bias has resulted from the presence of others (Castelli & Tomelleri, 2008), the presence of a Black compared to a White experimenter (Lowery, Hardin, & Sinclair, 2001), and endorsement of anti-racist sentiments by a (White) experimenter (Lun, Sinclair, Whitchurch, & Glenn, 2007; Sinclair, Lowery, Hardin, & Colangelo, 2005). In contrast, *increased* implicit bias has been observed under conditions in which participants are led to anticipate a public (versus private) discussion of bias scores (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Lambert et al., 2003) or following an interracial interaction (Amodio, 2009; Amodio & Hamilton, 2012).

In other studies, researchers have examined the influence of individual differences on internal factors, such as the motivation to be unbiased, on implicit bias. For example, several researchers have reported that participants who report less internal motivation to control their biases display stronger evaluative bias (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Gonsalkorale, Sherman, Allen, Klauer, & Amodio, 2011; Hausmann & Ryan, 2004) and more stereotypic bias (e.g., Amodio, Devine, & Harmon-Jones, 2008; Park, Glaser, & Knowles, 2008).

To date, no research has investigated the influence of both internal and external factors on measures of both evaluative and stereotypic racial bias in the same participants. Moreover, specific mechanisms for the influence of such moderating factors remain unclear. In some cases, evidence appears to suggest that such moderators affect bias by either reducing (Amodio, Harmon-Jones, & Devine, 2003; Gonsalkorale et al., 2011) or increasing (Amodio & Hamilton, 2012; Conrey et al., 2005) activation of automatic associations. Still other studies have suggested that moderators affect bias through reduced (Amodio et al., 2009; Conrey et al., 2005; Lambert et al., 2003) or increased (Amodio et al., 2008) engagement of control-related processes. Differences in task structure and content (and, therefore, the automatic and controlled processes they elicit) as well as differences in the way moderators are operationalized across studies make it difficult to resolve these inconsistencies. Thus, the current study investigated the effect of an observer (external) and internal motivation to be unbiased (internal) to provide a side-by-side comparison of the effects of external and internal factors on response bias in two different implicit bias tasks.

Internal motivation to be unbiased. The Internal (IMS) and External (EMS) Motivation to respond without prejudice Scales (Plant & Devine, 1998) were administered during a mass testing session several weeks prior to the experiment. The IMS consists of five items tapping a personal desire to be unprejudiced, such as, “Being nonprejudiced toward Black people is important to my self-concept.” Items were reverse-scored as necessary and averaged so that higher scores indicate greater internal motivation. The EMS also consists of five items, tapping normative reasons for appearing nonprejudiced, such as, “I attempt to appear nonprejudiced toward Black people in order

to avoid disapproval from others." For each item participants indicated their agreement using a 1 (*strongly agree*) to 9 (*strongly disagree*) Likert scale. Items were reverse-scored as necessary and averaged so that higher scores indicate greater external motivation. Internal consistency was acceptable in the current sample (IMS: $\alpha = .87$, EMS: $\alpha = .77$ in Study 1; IMS: $\alpha = .85$, EMS: $\alpha = .80$ in Study 2). Consistent with previous research (e.g., Devine et al. 2002), IMS and EMS scores were uncorrelated ($r = .12$, $p = .287$ in Study 1; $r = -.10$, $p = .153$ in Study 2).

A difference score (IMS-EMS) was also created to examine the influence of internal motivation, accounting for external motivation. Previous research suggests the degree to which motivation has been internalized is an important predictor of ability to control biased responses (Amodio, Harmon-Jones, and Devine (2003). IMS, EMS, and the IMS-EMS difference score were all standardized.

Motivation and observer condition as moderators. We separately examined the effects of IMS, EMS, the IMS-EMS difference score, and the presence of an observer (effect coded: Absent = -1, Present = 1) on PDP-C and PDP-A estimates in both tasks. Additionally, the interaction of each variable with Task (effect coded: APT = -1, WIT = 1) was examined. To accomplish this, we ran four models on PDP-A estimates, with one of the variables of interest, Task, and the interaction between the variable and Task included as predictors in each. Four identical models were also run on PDP-C estimates. In Study 1, the only significant effect that emerged was the effect of EMS on PDP-A estimates, $\beta = .22$, $p = .005$, such that higher external motivation predicted a larger contribution of automatic processing. This effect was not qualified by Task, $\beta = .04$, $p = .570$, suggesting that EMS predicted PDP-A in both tasks. This relationship replicated in

Study 2, $\beta = .15$, $p = .011$, and the interaction with Task was again non-significant, $\beta = .01$, $p = .840$.

Although the relationship between EMS and PDP-A estimates was the only significant relationship in Study 1, several other significant effects emerged in Study 2. The effect of IMS-EMS on PDP-A estimates was also significant, $\beta = -.16$, $p = .006$, and not qualified by task, $\beta = -.03$, $p = .538$, such that more internalization of motivation to be unbiased predicted a smaller contribution of automatic processing. Additionally, both IMS, $\beta = .16$, $p = .013$, and the IMS-EMS difference score, $\beta = .18$, $p = .008$, significantly predicted PDP-C scores. Neither was qualified by Task. These results suggest a greater degree of internalization predicts greater control exhibited in each task.

The lack of significant effect of the observer manipulation stands in contrast to previous studies that have shown significant moderation of bias as a result of external factors, including the presence of others, expression of egalitarian values by others, or anticipated public discussion of one's bias (e.g., Amodio, 2009; Amodio & Hamilton, 2012; Boysen, Vogel, & Madon, 2006; Castelli & Tomelleri, 2008; Lambert et al., 2003; Lowery et al., 2001; Sechrist & Stangor, 2001; Sinclair et al., 2005). Our observer manipulation was most similar to that employed in Castelli and Tomelleri (2008), where subjects completed a race IAT (Study 1) or a lexical decision task (Study 2) either alone or in the same room with two other subjects. The authors reported that the presence of others reduced anti-Black evaluative bias, primarily because participants in the "presence of others" condition slowed down their responses in bias-congruent trials, relative to participants in the "alone" condition. This tactic was not available in the current studies due to the fast response deadline, which may explain the lack of effect found in the

current studies. In other words, it is possible that despite the subtlety of the observer manipulation used in the current research, we may have seen an effect of an observer if the affective priming task had not employed a response deadline. Alternatively, both evaluative and stereotypic implicit bias may be more resistant to influence by external factors than recent research suggests (e.g., Sinclair, Kenrick, & Jacoby-Senghor, 2014). However, in the absence of evidence in support of the null hypothesis, assertions that the observer did not have an effect are not possible.

Effect of the Race of the Observer

In Study 1, experimenters included two White female, one White male, and one Asian/White female. In Study 2, experimenters included ten White males, nine White females, one Black male, one Black female, one Asian/White male, and one Asian/White female. Because we had a larger and more diverse group of experimenters in Study 2, we additionally examined the race of the observer on racial bias. Experimenter race was coded as "White" or "Non-white". A multilevel model was fitted to response accuracy bias scores for each task with experimenter race, task, and presence of observer included as predictors. The intercept was allowed to vary by subject. The main effect of experimenter race was nonsignificant, $b = .02$, $p = .563$. Additionally, no interactions were significant, ps $> .19$. Given the small number of non-White experimenters, we cannot conclusively state that the race of the experimenter did not have an effect on response accuracy bias. However, we do not have evidence that the race of the experimenter significantly impacted participants' response bias in either task.

Replacing Repeated-Measures ANOVA with MLM

Multilevel modeling (MLM) is an alternative approach to analyzing repeated measures ANOVA and allows for more flexible model specifications, including the use of continuous predictors. Although we use repeated measures ANOVA when appropriate in the main text, we provide here parallel analyses using MLM and show that the pattern is identical. MLM can account for missing data and does not require list-wise deletion. Thus, all participants with task data were included (n = 100 in Study 1, n = 204 in Study 2). The random effects structure for all models allowed covariance between random slopes and intercepts; in each model, as many slopes varied by subject as model convergence allowed. Satterthwaite approximations were used to estimate degrees of freedom and to obtain two-tailed *p*-values; in situations where the degrees of freedom exceeded 200, we report the results as *z* statistics. More details about the random effects structure used in each model can be found at https://www.github.com/hiv8r3/Observer_2studies in the file labeled "Analyses__SM_.html".

Study 1 results. Examination of error rates from the WIT using MLM showed the same pattern of results as repeated measures ANOVA. Most importantly, there was a significant Prime x Target interaction, $\beta = -.23$, $t(182.0) = -12.2$, $p < .001$, such that guns were categorized more accurately than tools following Black faces, whereas tools were categorized more accurately than guns following White faces. The Prime x Target interaction in the APT also matched the interaction revealed by repeated measures ANOVA, $\beta = .18$, $t(184.0) = 8.0$, $p < .001$, such that negative words were categorized more accurately than positive words following Black faces and positive words were categorized more accurately than negative words following White faces. Also identical to

the results reported in the main text, examination of error rates from both tasks revealed a significant Prime x Target x Task interaction, $\beta = .20$, $z = 6.2$, $p < .001$, indicating that patterns of race bias differed across the tasks.

Study 2 results. Examination of error rates from the WIT using MLM showed the same pattern of results as repeated measures of ANOVA. Most importantly, there was a significant Prime x Target interaction, $\beta = -.18$, $z = -13.6$, $p < .001$, such that guns were categorized more accurately than tools following Black faces, whereas tools were categorized more accurately than guns following White faces. The Prime x Target interaction in the APT also matched the interaction revealed by repeated measures ANOVA, $\beta = .15$, $z = 9.5$, $p < .001$, such that negative words were categorized more accurately than positive words following Black faces and positive words were categorized more accurately than negative words following White faces. Also identical to the results reported in the main text, examination of error rates from both tasks revealed a significant Prime x Target x Task interaction, $\beta = .24$, $z = 10.9$, $p < .001$, indicating that patterns of race bias differed across the tasks.

Supplementary References

Amodio, D. M. (2009). Intergroup anxiety effects on the control of racial stereotypes: A psychoneuroendocrine analysis. *Journal of Experimental Social Psychology*, *45*(1), 60–67. https://doi.org/10.1016/j.jesp.2008.08.009

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

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, *94*(1), 60–74. https://doi.org/10.1037/0022-3514.94.1.60

Amodio, D. M., & Hamilton, H. K. (2012). Intergroup anxiety effects on implicit racial evaluation and stereotyping. *Emotion*, *12*(6), 1273–1280. https://doi.org/10.1037/a0029016

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology*, *84*(4), 738–753. https://doi.org/10.1037/0022-3514.84.4.738

Boysen, G. A., Vogel, D. L., & Madon, S. (2006). A public versus private administration of the implicit association test. *European Journal of Social Psychology*, *36*(6), 845–856. https://doi.org/10.1002/ejsp.318

Castelli, L., & Tomelleri, S. (2008). Contextual effects on prejudiced attitudes: When the presence of others leads to more egalitarian responses. *Journal of Experimental Social Psychology*, *44*(3), 679–686. https://doi.org/10.1016/j.jesp.2007.04.006

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.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835–848. https://doi.org/10.1037/0022-3514.82.5.835

Gonsalkorale, K., Sherman, J. W., Allen, T. J., Klauer, K. C., & Amodio, D. M. (2011). Accounting for successful control of implicit racial bias: The roles of association activation, response monitoring, and overcoming bias. *Personality and Social Psychology Bulletin*, *37*(11), 1534–1545. https://doi.org/10.1177/0146167211414064

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

Hausmann, L. R. M., & Ryan, C. S. (2004). Effects of external and internal motivation to control prejudice on implicit prejudice: The mediating role of efforts to control

prejudiced responses. *Basic and Applied Social Psychology*, *26*(2–3), 215–225. https://doi.org/10.1080/01973533.2004.9646406

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

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*(5), 842–855. https://doi.org/10.1037/0022-3514.81.5.842

Lun, J., Sinclair, S., Whitchurch, E. R., & Glenn, C. (2007). (Why) do I think what you think? Epistemic social tuning and implicit prejudice. *Journal of Personality and Social Psychology*, *93*(6), 957–972. https://doi.org/10.1037/0022-3514.93.6.957

Park, S. H., Glaser, J., & Knowles, E. D. (2008). Implicit Motivation to Control Prejudice Moderates the Effect of Cognitive Depletion on Unintended Discrimination. *Social Cognition*, *26*(4), 401–419. https://doi.org/10.1521/soco.2008.26.4.401

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Sechrist, G. B., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. *Journal of Personality and Social Psychology*, *80*(4), 645–654. https://doi.org/10.1037/0022-3514.80.4.645

Sinclair, S., Kenrick, A. C., & Jacoby-Senghor, D. S. (2014). Whites' Interpersonal Interactions Shape, and Are Shaped by, Implicit Prejudice. *Policy Insights from*

the Behavioral and Brain Sciences, *1*(1), 81–87. https://doi.org/10.1177/2372732214549959

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*(4), 583–592. https://doi.org/10.1037/0022-3514.89.4.583

Figure S1. Associations between task-wide automatic (PDP-A) and controlled (PDP-C) processing estimates across the stereotypical (WIT) and evaluative (APT) bias tasks. PDP-A estimates represent the difference between White-prime *A* and Black-prime *A* trials in both tasks (Black – White). All PDP estimates are z-scored.